National Library of Canada
Bibliothèque nationale du Canada

Canadian Theses Service
Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

## NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

## AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

NL-339 (r. 88/04) c

Canada

UNIVERSITY OF ALBERTA

# PERCEPTION AND PRODUCTION OF ENGLISH VOWELS BY NATIVE SPEAKERS OF ARABIC

BY

MURRAY J. MUNRO

(C)

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

IN

SPEECH PRODUCTION AND PERCEPTION

DEPARTMENT OF LINGUISTICS

EDMONTON, ALBERTA
SPRING 1992

National Library of Canada
Bibliothèque nationale du Canada

Canadian Theses Service
Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-73043-9

Canada

UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: Murray J. Munro

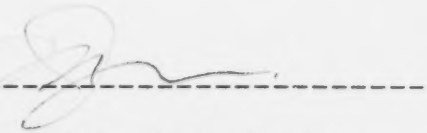
TITLE OF THESIS: Perception and Production of English Vowels by Native Speakers of Arabic

DEGREE: Ph.D.

YEAR THIS DEGREE GRANTED: 1992

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.



1300 - 20th Street South, Apt. #105
Birmingham, AL USA 35205

Date: Oct. 14, 1991

And it was so, that when those Ephraimites which were escaped said, Let me go over; that the men of Gilead said unto him, Art thou an Ephraimite? If he said, Nay; then said they unto him, Say now Shibboleth: and he said Sibboleth: for he could not frame to pronounce it right. Then they took him, and slew him at the passages of Jordan: and there fell at that time of the Ephraimites forty and two thousand.

Judges 12:5-6

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

THE UNDERSIGNED CERTIFY THEY HAVE READ, AND RECOMMEND TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH FOR ACCEPTANCE, A THESIS ENTITLED PERCEPTION AND PRODUCTION OF ENGLISH VOWELS BY NATIVE SPEAKERS OF ARABIC

SUBMITTED BY MURRAY J. MUNRO

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

IN SPEECH PRODUCTION AND PERCEPTION.

Terrence M. Nearey (Supervisor)

Bruce L. Derwing

Kyril T. Holden

Bernard L. Rochet

John J. Ohala

per Robert F. Port

Date:

This thesis is dedicated to my father, James J. Munro.

**Abstract**

This study consists of four experiments which explored the perception and production of vowels by native speakers of Arabic and native speakers of English. In the first study, identification data from a synthetic /bit/-/bIt/ continuum, in which duration and spectrum were varied independently, indicated that the Arabic listeners showed greater relative sensitivity to duration than did a group of native English listeners. This finding was supported by the results of another experiment in which the two groups of listeners were trained to identify synthetic exemplars of /y:/ and /ø/, and were then tested on /y/ and /ø:/ tokens. In the third experiment, measurements were obtained of the durations and formant frequencies of ten English vowels produced by the two groups in /bVt/ and /bVd/ contexts. On virtually every vowel the Arabic speakers' productions differed from those of the English speakers in some respect, such as duration, F1 or F2 frequency, or degree of F1 or F2 movement. Finally, a rating experiment was performed, in which five judges assessed the degree of accentedness of the front vowels produced in the production study. Acoustic data were regressed on the ratings in order to determine which properties of the vowels caused them to sound accented. Such factors as F1 frequency, F2 movement, and duration emerged as significant predictors of the ratings. Overall, the results of these experiments indicated that the Arabic speakers neither perceived nor produced the English vowels in a "native-like" manner, even though they had spoken English for an average of nearly eight years. This study also illustrates how accentedness in vowels may by quantified acoustically and related to perceptual data.

**Acknowledgements**

The list of those who have assisted in some way with this research is very long, and unfortunately not everyone can be named here. I would first like to express my appreciation to my supervisor, Terry Nearey, for his encouragement and help, not only during the preparation of this thesis, but throughout my graduate program. I am also grateful to the other committee members, Bruce Derwing, Kyril Holden, John Ohala, and Bernard Rochet, and to the external reader, Robert Port, for their helpful comments and suggestions. Special thanks are due to James Flege who served as "surrogate superviser" in Birmingham, and whose insights have greatly influenced this work.

For assistance in preparing stimuli, I would like to thank Paula Crosby, Katherine Fletcher, and Steve McKinney, all of whom listened to several annoying sequences of half-baked synthetic vowels. For technical assistance, thanks are due to Steve Smith in Birmingham, and to Tom Welz and Terry Baxter in Edmonton, who were always available to answer a question or assist with a program that didn't work correctly. I also greatly appreciate the help of nearly 50 subjects, both English- and Arabic-speaking, and miscellaneous assistance from Grace Wiebe and Sally Rice.

Financial support from a SSHRC Doctoral Fellowship, the Government of Alberta, the University of Alberta, NIH Grant #DC00257, and the Department of Linguistics is gratefully acknowledged. I would also like to thank the Department of Biocommunication at UAB for the use of its facilities.

It would be difficult to live a sane existence as a graduate student without the encouragement of friends. Here I would would like to thank three special people without whose friendship I would be much the poorer: my fellow student, Judy Cameron, for many hours of chatting and commiserating; my long-time colleague, Tracey Derwing, for being

a critical, yet always supportive listener; and my friend, Tommy Thompson, for always wanting to know what I was doing and why.

**Table of Contents**


Chapter 1: Introduction ..... 1

Speech Perception in Adults and Children ..... 2

Evidence from Second Language Acquisition ..... 9

Overview of the Present Research ..... 20

References ..... 21

Chapter 2: Spectral and Temporal Cues in Vowel Perception by Native Speakers of English and Arabic ..... 28

Introduction ..... 28

Experiment I: Identification of a synthetic English /bit/-/bIt/ continuum ..... 35

Methods ..... 35

Results ..... 38

Discussion ..... 44

Experiment II: Identification of two foreign vowel categories by speakers of Arabic and English ..... 50

Methods ..... 50

Results ..... 55

Discussion ..... 57

Tables and Figures for Chapter 2 ..... 61

References ..... 71

Chapter 3: Experiment III: English Vowel Production by Native Speakers of Arabic ..... 74

Introduction ..... 74

Methods ..... 79

Results ..... 81

Discussion ..... 91

Tables and Figures for Chapter 3 ..... 100

References ..... 111

Chapter 4: Experiment IV: Accentedness Judgments and Acoustic Properties of Foreign–Accented Vowels ..... 114
Introduction ..... 114
Methods ..... 119
Results ..... 121
Discussion ..... 127
Tables and Figures for Chapter 4 ..... 132
References ..... 135

Chapter 5: General Discussion and Conclusions ..... 137
Effects of L1 on L2 ..... 137
Individual Differences ..... 139
The Relationship between Perception and Production ..... 140
Measuring Accentedness ..... 143
References ..... 145

Appendices ..... 146
Appendix A: Instructions to Subjects ..... 146
Appendix B: Nominal and Measured Properties of Stimuli in Experiment I ..... 148
Appendix C: Nominal and Measured Properties of Stimuli in Experiment II (Hz) ..... 149

## List of Tables


Table 2-1: AR Subjects in Experiment I ..... 61
Table 2-2: Stimuli in Experiment I ..... 62
Table 2-3: % Change in 'Beat' ID at each Spectral Step as a Function of Duration ..... 62
Table 2-4: Comparison of Prediction Models (Experiment I) ..... 63
Table 2-5: Formant Values of Stimuli in Experiment II ..... 63
Table 2-6: Summary of 'A' and 'B' (%) Scores in Experiment II ..... 63
Table 2-7: Results of Experiment II by Subject (% Duration-matches on 'B' Stimuli) ..... 64
Table 3-1: Summary of Vowel Inventories by Dialect ..... 100
Table 3-2: Mean Durations by Vowel (ms) ..... 100
Table 3-3: Mean Durations by Final Consonant (ms) ..... 100
Table 3-4: Mean Formant Frequencies (Hz) ..... 101
Table 3-5: Mean change in F1 and F2 (Hz) ..... 102
Table 3-6: Comparison of Log-transformed Formant Values from the Two Groups (two-tailed t - values) ..... 103
Table 3-7: Comparison of Individual Duration Data and Perception Scores from Experiment I ..... 104
Table 3-8: Individual Differences from English Means on F1 and F2 of /i/ and /I/ (Hz) ..... 105
Table 3-9: Individual Spectral Differences on Three Vowels ..... 106
Table 3-10: Correlations between Perceptual Data (Experiment I) and Spectral Properties ..... 106
Table 4-1: Mean Ratings in Experiment IV by Speaker and Vowel ..... 132
Table 4-2: Inter-rater Agreement in Experiment IV (Pearson r) ..... 132
Table 4-3: Predictor Variables in Regression Analysis I ..... 133
Table 4-4: Summary of R2 Values from Two Regression Analyses ..... 133
Table 4-5: Significant Predictors of Judges' Ratings in the Two Analyses ..... 133

## List of Figures

Figure 2-1: Identification Scores by Spectral Step ..... 65
Figure 2-2: Identification Scores by Duration Step ..... 66
Figure 2-3: Effect of Spectral Properties at Each Duration Step ..... 67
Figure 2-4: Territorial Maps ..... 68
Figure 2-5: Change in % 'Beat' ID Due to Spectral and Temporal Properties ..... 68
Figure 2-6: Spectral vs. Temporal Coefficients from the Linear Logistic Analysis ..... 69
Figure 2-7: Stimuli in Experiment II ..... 70
Figure 3-1: Durations of Front Vowels ..... 107
Figure 3-2: Durations of Back Vowels ..... 107
Figure 3-3: Vowel Duration by Final Consonant ..... 108
Figure 3-4: Mean Formant Values (Measurement A) ..... 108
Figure 3-5: Change in F1 ..... 109
Figure 3-6: Change in F2 ..... 110
Figure 4-1: Hypothetical Rating Function ..... 134

# CHAPTER 1
# INTRODUCTION

It has long been observed that those who learn a second language (L2), particularly in adulthood, often fail to achieve native–like pronunciation and, despite years of experience, always exhibit some degree of "foreign accent." One of the central problems facing researchers seeking to understand how the sound system of a second language is acquired is to explain the causes of this phenomenon. Here many difficult questions arise. Why do adults acquire an L2 sound system less well than children? Are L2 learners' errors predictable, and if so, on what basis? How can foreign accent be quantified? To what extent are the production errors of L2 learners the result of 'errors' in perception?

If we are to understand why a foreign accent occurs, we must first be able to describe what it is. Yet it is only quite recently that researchers have begun to deal with L2 production in the fine–grained way that the phonetics laboratory makes possible. It may be useful to observe that native Mandarin speakers of English have difficulty with the production of final stop voicing, but if we are to understand such a phenomenon fully, we need to know as much as possible about the phonetic details of the speakers' productions, especially with respect to properties known to be of perceptual relevance to native speakers of L2. Are release bursts produced, and, if so, how intense are they? Are proper vowel duration differences before voiced and voiceless articulations maintained? How much voicing occurs during stop closure? Does F1 show voicing-conditioned differences in offset frequency? Questions such as these cannot be answered with precision without an instrumental analysis of L2 learners' productions. Once detailed data of this sort have been gathered, they can be compared with measurements of the productions of native speakers, and conclusions can be drawn about

precisely what characterizes the L2 learners' patterns of errors. Perhaps just as important as studies of production data is a comparison of perceptual data from native speakers and L2 learners. Discrimination and identification tasks with synthetic and natural stimuli might reveal perceptual 'errors' which underlie non–native production patterns.

While much research has been conducted on the production and perception of L2 speech sounds, there are still many important questions which remain unanswered. The review presented below discusses some of the key research in this area and highlights some of the major issues.

## Speech Perception in Adults and Children

### *Differences among Infants, Children, and Adults*

One cannot attempt to address the question of why adults usually exhibit imperfect learning of L2 speech sounds without giving some attention to the growing body of research on the development of speech perception and production from infancy onward. It is essential to consider what perceptual and productive abilities exist at early stages of human development and how those abilities change over time.

Using a high–amplitude–sucking (HAS) paradigm, Eimas, Siqueland, Jusczyk, & Vigorito (1971) demonstrated not only that infants could discriminate voiced and voiceless stop consonants, but also that their within–category discrimination was poorer than their between–category discrimination. The suggestion that speech perception in infants is categorical has provoked considerable controversy over whether or not infants are innately endowed with the ability to perceive human speech in a phonetic mode. Since that time, much attention has been given to the ability of infants to perceive phonetic contrasts, and it is generally accepted that infants

can indeed discriminate almost every contrast on which they have been tested. (See Kuhl, 1987 for a detailed review.)

At the same time, however, there is considerable research showing that without training, adults often perform poorly on discrimination tasks involving contrasts with which they are unfamiliar (see Burnham, Earnshaw, & Quinn, 1987). Miyawaki, Strange, Verbrugge, Liberman, Jenkins, & Fujimura (1975), for instance, found that native Japanese speakers discriminated a synthetic English /rɑ/–/lɑ/ continuum very poorly and showed no better performance between categories than within. Werker, Gilbert, Humphrey, & Tees (1981) and Werker & Tees (1983) performed perceptual experiments involving the Hindi dental–retroflex /t̪ɑ/–/ʈɑ/ contrast, using multiple exemplars of natural tokens in a simple discrimination paradigm. While six– to eight–month–old infants (tested using the head–turn procedure) and Hindi–speaking adults reached criterion quite readily, native English adults performed significantly less well. Moreover, when children at 4, 8, and 12 years of age were tested, they too performed poorly. Training with feedback improved the adults' discrimination scores little, although it did help somewhat on the Hindi /tʰɑ/–/dʰɑ/ contrast which was also discriminated less well by the adults than by the infants. A similar inability to discriminate was found with synthetic stimuli representing the same contrast (Werker & Lalonde, 1988) and with natural tokens representing the Thompson glottalized velar and glottalized uvular contrast (Werker & Tees, 1984a).

The results of studies such as these argue for early changes in the way speech is perceived. Werker & Tees (1984b) and Werker & Lalonde (1988) found evidence that such a change occurs or begins to occur at some point during the first year of life, since 6–8 month old infants performed significantly better than 11–13-month-old infants on the Thompson and Hindi place contrasts. Evidence for very early effects of linguistic experience also comes from production studies.

It has been demonstrated, for example, that as early as six months, infants show the influence of their linguistic environment in the formant frequencies of their vocalizations (de Boysson–Bardies, Halle, Sagart, & Durand, 1989).

As Aslin and Pisoni (1980) have indicated, the role of early experience in speech perception is likely to be a complex one. First, it is conceivable that certain contrasts are perceptible at birth and may be either facilitated or lost as a result of specific linguistic experience. Second, those contrasts which infants are incapable of perceiving (see Barton, 1980) may be induced through experience. But even for contrasts which are discriminable at birth, several years of experience may be required before perception and production are completely adult-like. Strange & Broen (1980), for example, found differences between adults and 3–year–olds in boundary locations for /l/–/r/ and /w/–/r/. Also Bernstein (1983) found that 4– and 6–year–old children did not show the trading relation between F0 and VOT which is known to occur in adults. Of course, a final possibility is that the perception of some contrasts may remain largely unaffected by linguistic experience. Best, McRoberts, & Sithole (1988) found that English-speaking adults had no difficulty perceiving contrasts among Zulu clicks to which they had had no prior exposure.

The research by Werker and her colleagues has often been taken to mean that ability to discriminate certain speech contrasts is "lost" at some early stage of development. This view requires some further elaboration. First, given the right testing conditions or training procedures, adult discrimination scores may improve considerably. Werker & Tees (1984a) found that English–speaking adults were able to perform at levels above chance on both the Hindi and Thompson contrasts, provided the interstimulus interval was reduced to 500 ms. (In other experiments, the ISI was 1500 ms or greater.) They used this finding to argue that under short ISI

conditions their listeners had used a level of processing intermediate between a linguistic level (i.e., a level at which perception would occur in terms of the native language sound system) and an auditory one. Since the subjects heard multiple natural tokens, there was some within–category variability in the stimulus items, though just how much variability was involved is not clear. The fact that subjects were able to recognize that certain pairs of sounds belonged to the same phonetic category even though they were not physically identical suggests that a purely auditory strategy was not being used. On the other hand, a purely linguistic strategy would have led them to fail to discriminate between the two categories entirely (as was the case when the ISI was longer), because English does not make the place distinctions being tested. Werker and Tees concluded that a relatively short ISI allowed the listeners to "relinquish a purely phonemic strategy" in making their judgments.

Whether or not certain contrasts are truly lost, the facts nevertheless indicate differences in speech perception between adults and children. One account of the results discussed above and other experimental findings has been proposed by Burnham (1986) and Burnham, Earnshaw, & Quinn (1987). Drawing upon evidence from a number of cross–language infant speech perception studies, they have argued that segment–level contrasts in languages may be relatively "robust" or "fragile." Robust contrasts are those which have a fairly strong psychoacoustic basis and which are likely to exist in a large number of languages. Fragile contrasts, on the other hand, have less psychoacoustic salience and are less common cross–linguistically. Burnham (1986) argues that the ability to perceive fragile contrasts, such as the glottalized velar–uvular contrast in Thompson (Werker et al., 1984) or the /r/ – /w/ contrast in English are lost relatively early, perhaps in the first 6 -12 months of infancy, unless the linguistic environment requires that the contrast

be maintained, while the perception of robust contrasts persists until 4-8 years and is easier to recover in adulthood.

This proposal undoubtedly deserves further exploration. However, there are some problems with it. Specifically, there is no clear criterion which can be applied a priori to determine which contrasts are perceptually robust. Therefore, it is very difficult to test the claim that a certain contrast is readily learned by adults *because* it is a robust contrast. The mere observation that a certain pair of sounds occurs rarely in phonetic inventories does not constitute proof that it is not a perceptually robust pair. As Best et al. (1988) observed, clicks are relatively rare sounds, but adult English listeners find at least some pairs of clicks quite easy to discriminate. It is conceivable that the scarcity of these or other sounds in linguistic inventories may be explained as a consequence of production difficulty. The lack of a means of determining which contrasts are robust results in circularity in the Burnham's hypothesis. While it *predicts* that robust contrasts will occur in a wide range of languages and are relatively easy to recover, he uses the same criteria to *define* "robust contrast." Independent grounds would have to be established for determining robustness if the proposal were to be tested seriously. One might look at those contrasts which appear to pose differing degrees of difficulty for infants, but one can not be sure whether the difficulties arise from properties of the speech sounds which make them hard to discriminate or from underdeveloped perceptual mechanisms in the infant. Another possibility is to assess the relative salience of contrasts to non–human subjects. Some recent research of this type has proved encouraging. Sinott (1989), for instance, has shown that English vowel contrasts which are thought to be relatively difficult for humans are also difficult for monkeys.

Best (1990) has proposed that the difficulty adults have in perceiving particular non–native contrasts is related to how the foreign categories are assimilated to native ones. She presents

evidence for four types of assimilation: *two–category* , in which two non–native phones are assimilated to two different native categories; *category goodness* , in which both phones are perceived as belonging to one native category, though one is a good exemplar and the other is not; *single category* , in which both phones are perceived as poor exemplars of one category; and *non–assimilable* , in which neither phone is perceived as belonging to a native category. For the first type, excellent discrimination by adults is predicted, while for the second and fourth types good discrimination is possible. In the case of single category assimilation, however, poor discrimination is predicted. Her evidence is based on experiments with monolingual English speakers tested in an AXB discrimination paradigm on a variety of foreign contrasts.

***Other Training Experiments***

The studies discussed above leave open the question of whether the relatively poor performance of adults on non-native contrasts was due to the training methods used. It has been suggested that with the right procedures and possibly with the right amount of input it may be possible to direct the attention of adult subjects to the critical properties of the stimuli and therefore improve their performance. Strange & Dittmann (1984) had only very limited success in improving perceptual performance on an English contrast in the laboratory. Although they were able to improve identification and discrimination scores on /r/ and /l/ among Japanese adults using a discrimination task with feedback, the improved performance did not transfer to natural speech. However, other research suggests that the degree of success and transferability of learning may depend on the type of training task used. Logan, Lively, & Pisoni (1991), for instance, reported more success in an identification task, again involving Japanese speakers, which emphasized variability among

exemplars of /r/ and /l/. Natural tokens from different speakers were used in the training phase.

Pisoni, Aslin, Perey, & Hennesy (1982) and McClaskey, Pisoni, & Carrel (1983) were successful in training English speakers to classify a VOT continuum into three categories rather than two by using a procedure in which subjects were trained on good synthetic exemplars of each category and then asked to categorize other tokens with feedback. Most, though not all, subjects reached criterion (85% correct) after about 1.5 hours of training. Moreover, training on stimuli representing one place of articulation readily transferred to other places with no additional training. Flege & Wang (1990) were able to improve significantly the perception of the English final /t/-/d/ contrast among speakers of three Chinese languages in an identification task which used natural tokens with release bursts and voicing removed. Jamieson and Morosan (1986) used a perceptual fading technique with feedback to train speakers of Canadian French to attend to the English /θ/ - /ð/ contrast. A maximum of 90 minutes of training over 4 sessions was given using synthetic stimuli varying in duration of frication. After training, identification scores improved significantly, even on natural tokens which had not been heard during the training phase of the experiment. In attempting to account for the differences in success between their method and experiments involving discrimination training, Jamieson and Morosan suggest that discrimination tasks may cause listeners to attend to differences between stimuli, rather than focus on shared properties of tokens belonging to a category. Ultimately, then, discrimination tasks may be less successful in the training of new contrasts. In a further study, Morosan & Jamieson (1989) showed that training effects using the same procedure transferred to a variety of new, natural voices not used in the training phase, but not to new phonetic contexts.

It is important to note that in the training studies discussed above, the subjects did not reach 100% identification or discrimination scores. There are several possible explanations for this. It is conceivable that not enough training was provided and that further training would have eventually led to native-like performance. Another possibility is that the "perfect" training paradigm has not yet been developed, but that with further research, it will eventually be found. It might be argued, however, that the laboratory environment is simply too contrived to elicit the best possible performance from listeners. Laboratory training procedures are likely to be at best rather artificial. For practical reasons, the experimenter may fail to use the range of stimuli and speakers likely to be encountered in natural communicative situations; listeners may not respond to tape recorders, headphones, and computer terminals as they would to more "natural" learning conditions; and they may simply not be motivated to do their best. Obviously, an experimenter cannot hope to provide to subjects the quantity and quality of input which normal L1 learners would receive (e.g. over the first several years of childhood). There is good reason, then, to have modest expectations about what can be achieved in the laboratory.

## Evidence from Second Language Acquisition

### *Limits on Adult Performance*

Of course another possible explanation for the lack of native-like performance in the training studies is that some upper limit on performance by adult subjects was reached. That is, for some reason or reasons, adults are not normally able to learn new contrasts perfectly. In fact, there is now a substantial body of evidence indicating that adults who learn a second language do not generally show perceptual and productive mastery of the L2 sound system, even after speaking their second language for many years.

Perceptual studies involving speakers of a variety languages and several different speech sounds have shown that L2 learners' perceptions often differ noticeably from those of native speakers (e.g., Flege & Eefting, 1987; Gottfried & Beddor, 1988; Flege & Bohn, 1989; Logan, Lively, & Pisoni, 1991). In addition, many studies have found comparable production differences (e.g., Mitleb, 1981; Flege & Port 1981; Port & Mitleb, 1983; Flege & Eefting, 1987).

A number of researchers have examined the role of certain extra-linguistic variables, particularly age, in success in L2 pronunciation. Tahta, Wood, & Loewenthal (1981) studied pronunciation scores assigned by three judges to 109 subjects who had learned English as a second language and who represented many linguistic backgrounds. A high negative correlation was found between the age at which L2 acquisition began and the degree of accent. Also important were whether English was spoken at home, age when tested, and sex. This study appears to confirm the widely-accepted view that older L2 learners are less likely than younger learners to develop native–like pronunciation (Oyama, 1982; Snow, 1987; Scovel, 1988). This view has not been uncontroversial, however, and in some respects the available evidence appears to be contradictory. For instance, Olson & Samuels (1982) found that college and junior high school students performed better than elementary school children on a foreign language pronunciation task, whereas Cochrane (1980) found that Japanese children received better English pronunciation scores than Japanese adults when rated by a panel of native speakers of English. Flege (1988) observed that foreign accents were detectable by native speakers of English even in a group of Taiwanese who had begun learning English in the United States at a mean age of 7.6 years. There is evidence then, that children do not always learn to pronounce foreign languages without an accent.

The different results in these studies can be at least partially accounted for if one considers that not all pronunciation studies have focussed on the same population. Olson & Samuels examined Americans who received 13 training sessions in German pronunciation over 3 weeks while in the United States. Tahta et al. and Cochrane, on the other hand, studied speakers of several languages who had been immersed in the L2 (American English) environment for some time. It appears that in the long run, after extensive exposure to the L2, younger learners are more apt to have native–like pronunciation, whereas adults may be able to learn pronunciation rules more quickly in a structured classroom setting, such as the one used by Olson and Samuels.

The Flege (1988) finding appears surprising because it has often been assumed that there is a relatively sudden decline in language–learning ability, including the ability to learn L2 pronunciation, in late childhood. It was once believed that such a decline occurred at puberty (Lenneberg, 1967), and was correlated with brain lateralization. However, this view has been largely discredited because lateralization is now believed to occur long before puberty (Krashen, 1973). The Flege study casts further doubt on the view that at some time in human development there is a sudden, dramatic loss of ability. Also, the finding of a correlation between age and degree of accent by Cochrane (1980) supports the view that any such change is gradual rather than abrupt. Taken together with the findings of Werker and her colleagues (see above), who observed changes in perception of foreign–language sounds even during the first year of life, these studies argue for a decreased sensitivity to non–native sounds which co–occurs with the learning of the L1 sound system and which somehow interferes with native-like acquisition of the L2 sound system.

Other studies indicate that accentedness in L2 productions is attributable to a variety of characteristics which distinguish L2

learners from L1 learners, such as age, degree of motivation, and even psychosocial factors, as well as the fact that L2 learners already have knowledge of the sound system of their native language. Suter (1976) and Purcell & Suter (1980), for instance, used a multiple regression analysis to determine which of a large set of variables related to personality and L2 experience best predicted the pronunciation scores assigned by a panel of 14 judges to 61 non–native speakers of English. Four predictors accounted for significant proportions of the variance in their scores: first language, aptitude for oral mimicry, length of residence in an English–speaking environment, and strength of concern for pronunciation accuracy. Some care must be taken in interpreting the highly significant correlation of the first of these predictors with the subjects' scores. It may be that speakers of some languages find it easier to learn to pronounce English than speakers of other languages, possibly because of the relationship between their native-language phonetic repertoires and the English inventory. On the other hand, it is conceivable that cultural differences between speakers of different languages may influence their motivation to pronounce a second language in a native–like manner. A further consideration is that judges may rate some types of accents more negatively than others simply because of biases against those accents, rather than on the basis of intelligibility, for instance, or some other criterion which could be applied equally to the productions of speakers of all languages.

***Effects of L1 on L2***

Judging from the results of Tahta et al., and Purcell & Suter, a number of factors probably underlie the difficulties faced by adult learners in acquiring the L2 sound system. Much of the research in this area has focussed on the ways that knowledge of the L1 system affects or "interferes" with L2 acquisition. It is frequently observed

that errors in L2 production show fairly direct influences of L1. For example, in studies of Jordanians and Saudis, Mitleb (1981), Flege & Port (1981) and Port & Mitleb (1983) observed the transfer of a number of properties of Arabic to English productions. Effects were seen in vowel durations, stop closure durations, and amount of aspiration in voiceless stops. When Flege & Eefting (1986) had Spanish-speaking learners of English categorize a VOT continuum, they found that the subjects' category boundaries were at significantly earlier times than those of monolingual English speakers; in other words, the boundaries were Spanish–like. Flege & Wang (1990) found evidence that the sensitivity to word final voicing contrasts in English among speakers of Mandarin, Shanghainese, and Cantonese was predictable on the basis of the degree to which the L1 permitted word-final consonants.

While it is often argued that production errors in L2 are the result of errors in perception, resulting from interference from L1, it has also been argued that a source of foreign accent is the transfer to L2 of voice quality settings appropriate for L1 (Esling & Wong, 1983). Such settings are defined as the "long–term postures" of the organs of speech which typify a particular language. For instance, a tendency toward tongue retroflexion (several languages of India), uvularization (Hebrew), pharyngealization (Arabic) or breathy voice may be maintained during segmental articulations of the new language, giving the L2 learner's productions the distinctive foreign accent associated with L1. In such a case, a foreign accent might be evident whether or not L2 perception was native–like. While this proposal seems quite plausible, it has yet to receive substantial empirical support.

***Predicting Pronunciation Errors***

Given that knowledge of the sound system of L1 adversely affects the learning of the L2 system, a logical question is whether the

errors made by L2 learners are readily predictable on the basis of linguistic analyses. Much of the early work on second language pronunciation was based on contrastive analysis, which received a great deal of attention during the 1950s and 1960s from linguists and those interested in second language pedagogy. This approach, which had its origins in structuralist linguistics, assumed that knowledge of a first language can "interfere" with the later learning of a second language. In general it was argued that in all aspects of second language acquisition, difficulties would be encountered by learners whenever there were differences between L1 and L2. The greater the difference between the two languages, the more difficulty the L2 learner would experience (Ellis, 1985). With respect to pronunciation, it was believed that an analysis of the phonological inventories of the first and second languages should make it possible to predict the types of errors made by L2 learners because they would tend to perceive and produce the sounds of the target language in terms of the L1 sound system. Lado (1957) argued that the easiest sounds to learn should be those which are physically close to the sounds in the L1. Because of 'positive transfer,' L2 learners should find it easy to learn phones which are similar to those in their first language, but 'interference' ought to make learning of radically different phones more difficult. Weinreich (1968) proposed various ways in which such interference might occur: reinterpretations, substitutions, underdifferentiations, and overdifferentiations. In addition, attempts were made to predict the degree of difficulty which the various types of interference would create for the L2 learner.

While there can be little doubt that native-language experience influences second language learning, many critics of the contrastive analysis (CA) approach argued that it failed to account for many of the errors made by learners. It was observed that L2 learners often had difficulty producing sounds which were regarded as similar to sounds in L1, and that often completely "new" sounds posed relatively

little difficulty (see Brown, 1980, for a review). Brière (1966), for instance, pointed out that analyses based entirely on phonemic descriptions often failed to account for data on L2 learners' errors. In a training experiment, he found unexpected asymmetries in success among English speakers who were taught several unfamiliar sounds from Vietnamese, French, and Arabic. While the subjects learned to produce /y/ and voiceless uvular /χ/ quite readily, according to the judgments of a panel of native speakers of languages using those sounds, they had much more difficulty with back unrounded /ɯ/ and voiced uvular /ʁ/. Two important problems are raised by these findings. First, the simple claim that whatever is "different" in the L2 system will pose difficulty for the learner is not upheld. Since /χ/ does not occur in English, we might incorrectly predict that English speakers will have a great deal of difficulty learning to produce it. Second, there is no obvious characteristic which emerges from a simple "feature–based" analysis of the English sound system which would lead one to predict the differing degrees of difficulty in the corresponding pairs Brière examined. In fact, one might argue that /y/ and /ɯ/ should be of equal difficulty because production of the former requires only the addition of a feature of roundedness to the English category /i/, while production of /ɯ/ requires only the removal of the same feature from /u/. Furthermore, contrastive analysis does not appear to account for the observation that English speakers do not learn /ʁ/ just as readily as /χ/, since once they have learned to articulate one of these sounds, they presumably need only transfer their knowledge of "voicing" to a new place of articulation to be able to produce the other.

Other researchers have reported findings which appear to show that positive transfer does not necessarily occur when it might be expected. Jamieson and Morosan (1986) observed that Canadian French speakers have difficulty perceiving the voicing contrast

between /θ/ and /ð/. Since French does not have interdentals, one might predict that these sounds will pose difficulty, at least in production. However, it does have a number of fricative pairs which are distinguished by the "voicing" feature (/s/-/z/, /f/-/v/, /ʃ/-/ʒ/). If a feature-based analysis provides all the necessary information to make a prediction, speakers of French should readily transfer their knowledge of fricative voicing to a new place of articulation in English, and should easily perceive the difference between voiced and voiceless interdentals. However, when subjects were asked to identify both natural and synthetic tokens as either /θ/ or /ð/ prior to training, they showed poor (chance level) performance.

Also, in several studies, Flege and his colleagues have shown that L2 learners often do not produce and perceive L2 speech sounds in a native-like way when they are similar to L1 sounds (Flege, 1984, 1987, 1988, 1991; Flege & Hillenbrand, 1984). He has suggested that this is because L2 learners do not take note of small phonetic differences between the native and L2 categories. Rather, because of the phenomenon of equivalence classification, similar L2 sounds tend to be perceived and produced in terms of L1 categories.

Brière (1966) concluded that "any prediction of a hierarchy of difficulty of learning phonological categories must be based on descriptions of these categories in terms of exhaustive information at the phonetic level, rather than on descriptions solely in terms of distinctive features or allophonic memberships of the phoneme classes" (p. 795). While this position is almost certainly correct, it should also be mentioned that certain information is essential to the testing of such predictions. In particular, discussions of L2 learners' errors have often relied far too much on anecdotal and often superficial information about the nature of those errors. Eckman (1977), for instance, presents a detailed explanation of certain L2 learners' errors in terms of markedness. Unfortunately, however,

much of his argument is based on anecdotal claims about difficulties faced by English speakers learning French and German speakers learning English. Here it seems reasonable to insist on a careful assessment of L2 productions before conclusions are drawn about the mechanisms underlying pronunciation errors. Both instrumental analyses and goodness judgments from native-speaking judges may be used to determine what, in fact, is difficult and what is not. Chapters 3 and 4 of the present study demonstrate how this may be achieved.

***One System or Two?***

A fundamental question in the acquisition of the L2 sound system is whether speakers of two languages are capable of learning two independent sound systems or whether they simply treat all the sounds from L1 and L2 as part of a single system. Caramazza, Yeni–Komshian, Zurif, & Carbone (1973) found evidence for the latter proposal in an identification experiment which showed that bilingual Canadian French-English speakers tended to have 50% crossover values at intermediate values of VOT which differed from VOT values for monolingual speakers of either language. They failed to obtain a "linguistic set" effect when they tested the subjects under two conditions in which the instructions were given in one of the languages. They observed comparable results in a production experiment as well. Obler (1982) presents similar data for English-Hebrew bilinguals, as does Williams (1977, 1979) for Spanish-English speakers. These studies considered "balanced" bilinguals - those who appear to be about equally competent in both languages, although one language is usually dominant. Although there were only small differences between the results in the two language conditions, the subjects performed differently from monolinguals who spoke either of their two languages.

On the other hand, Elman, Diehl, & Buchwald (1977) appeared to find a set effect in at least some of their Spanish-English bilingual subjects in a perceptual experiment involving natural bilabial tokens with varying values of VOT. They argued that differences in the method of testing were responsible for the contradictory results. In the other studies the experimenters attempted to induce the effect by keeping experimental materials, instructions, and the conversation preceding each of two separate tests (one for each language) in the language of interest. In the Elman et al. study, the test words were presented in either an English or a Spanish sentence frame. In general it was found that their most strongly bilingual subjects shifted their boundaries dramatically and performed like monolinguals, depending upon the context in which they heard a particular test item. One noteworthy result of the Elman et al. study was the finding that less-fluent bilinguals did not show the effect, while fluent bilinguals did. While their criteria for "fluency" are not entirely clear (it was assessed subjectively by one of the experimenters), this finding suggests that the ability to separate sounds from two languages is a function of how successfully the second language has been learned. An additional problem in interpreting the other studies is that they do not present data for individual subjects. Since there is a strong possibility of individual differences on these tasks, the presentation of only mean values from a group of subjects can obscure important facts about the data.

The view that a single sound system is used by L2 learners appears to be consistent with Flege's Speech Learning Model (SLM) (Flege 1987, 1988). He has argued that imperfect learning of the L2 inventory by adult L2 learners may be due to equivalence classification. As native–language categories are established, children learn to ignore certain within–category variations, such as small, non-phonemic differences in voice onset time, vowel formant targets, and fundamental frequency. Instead they focus on the

shared properties of a number of slightly different exemplars of particular categories. This mechanism may operate to prevent L2 learners from perceiving and producing many of the phones in the new system in a native–like manner. The phones likely to be learned less well are those which tend to be perceived as variants of already-existing L1 categories ("similar" phones). On the other hand, certain "new" sounds from the L2 may not be readily associated with L1 categories, and learners will eventually produce them correctly. In support of this account, Flege (1987) presented evidence that native English speakers of French produce the "new" French vowel /y/ more accurately than "similar" French /u/. He also compared the consonant productions of English speakers with varying degrees of experience in French to those of French monolinguals and of French speakers who had lived for several years in the United States. While the English speakers tended to approach the French VOT values as a function of experience, they also tended to reduce the VOTs in their English /t/ productions. Thus, it appears that learning a second language can influence L1 production. In this case, it might be argued that the reason for the effect on L1 was that the English speakers had not really established a new French /t/ category, but instead used a single /t/ category for both French and English. Other evidence for L2 effects on L1 comes from Garnes (1977), who observed that native speakers of Icelandic who had studied English outside Iceland perceived duration contrasts in their native language somewhat differently from monolinguals.

One problem with Flege's proposal, however, is that no clear criteria exist which can be used to determine which sounds from L2 are new and which are similar to sounds in the native language. A further problem concerns the way in which success in producing an L2 sound is assessed. Flege (1987) used F2 measurements from /y/ and /u/ to examine differences between the native French and native English groups. It might be argued that other properties of these

vowels should have been considered as well (e.g., F1, F3, and movement in F1 and F2).

**Overview of the Present Research**

The purpose of this study is to examine, from a variety of perspectives, some of the ways in which first language experience influences the learning of the sound system of a second language. First, an identification task is used to explore the use of spectral and temporal cues in the perception of English /i/ and /ɪ/ by native speakers of Arabic. In the second study, productions of ten English vowels in two consonantal contexts are considered. Finally, the relationship between certain acoustic properties of a set of accented English front vowels and accentedness judgments of native speakers is examined. The results of these studies are discussed in terms of how they relate to some of the issues discussed above.

**References**

Aslin, R., & Pisoni, D. (1980). Some developmental precesses in speech perception. In G. Yeni-Komshian, J. Kavanagh, & C. Ferguson (Eds.) *Child phonology, Vol. 2: Perception*. New York: Academic Press.

Barton, D. (1980). Phonemic perception in children. In G. Yeni-Komshian, J. Kavanagh, & C. Ferguson (Eds.) *Child phonology, Vol 2, Perception.* New York: Academic Press.

Beddor, P., & Strange, W. (1982). Cross-language study of perception of the oral–nasal distinction. *Journal of the Acoustical Society of America* **71**: 1551-1561.

Bernstein, L. (1983). Perceptual development for labelling words varying in voice onset time and fundamental frequency. *Journal of Phonetics* **11**: 383-393.

Best, C. (1990). *Adult perception of nonnative contrasts differing in assimilation to native phonological categories.* Paper presented at the meeting of the Acoustical Society of America, San Diego, CA.

Best, C., McRoberts, G., & Sithole, N. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English–speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance* **14**: 345-360.

Brière, E. (1966). An investigation of phonological interference. *Language* **42**: 768-796.

Brown, H. D. (1980). *Principles of language learning and teaching*. Englewood Cliffs, NJ: Prentice–Hall.

Burnham, D. (1986). Developmental loss of speech perception: exposure to and experience with a first language. *Applied Psycholinguistics* **7**: 207-240.

Burnham, D., Earnshaw, L, & Quinn, M. (1987). The development of the categorical identification of speech. In B. E. McKenzie, & R. H. Day (Eds.), *Perceptual development in early infancy.* Pp 237–275. Hillsdale N.J.: Lawrence Erlbaum Associates.

Caramazza, A., Yeni–Komshian, G., Zurif, E., & Carbone, E. (1973). The acquisition of a new phonological contrast: the case of stop consonants in French-English bilinguals. *Journal of the Acoustical Society of America* **54**: 421-428.

Cochrane, R. (1980). The acquisition of /r/ and /l/ by Japanese children and adults learning English as a second language. *Journal of Multilingual and Multicultural Development* **1**: 331-360.

de Boysson–Bardies, B., Halle, P., Sagart, L., & Durand, C. (1989). A crosslinguistic investigation of vowel formants in babbling. *Journal of Child Language* **16**: 1-17.

Eckman, F. (1977). Markedness and the contrastive analysis hypothesis. *Language Learning* **27**: 315-330.

Eimas, P., Siqueland, E., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science* **171**: 303–306.

Elman, J., Diehl, R., & Buchwald, S. (1977). Perceptual switching in bilinguals. *Journal of the Acoustical Society of America* **62**: 971-974.

Ellis, R. (1985). *Understanding second language acquisition.* Oxford: Oxford University Press.

Esling, J., & Wong, R. (1983). Voice quality settings and the teaching of pronunciation. *TESOL Quarterly* **17**: 89-95.

Flege, J. (1980). Phonetic Approximation in second language acquisition. *Language Learning* **30**: 117-134.

Flege, J. (1981). The phonological basis of foreign accent: a hypothesis. *TESOL Quarterly* **15**: 443-455.

Flege, J. (1984). The detection of French accent by American listeners. *Journal of the Acoustical Society of America* **76**: 692-707.

Flege, J. (1987). The production of "new" and "similar" phones in a foreign language: evidence for the effect of equivalence classification. *Journal of Phonetics* **15**: 47-65.

Flege, J. (1988). Factors affecting degree of perceived foreign accent in English sentences. *Journal of the Acoustical Society of America* **84**: 70-79.

Flege, J. (1991). Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language. *Journal of the Acoustical Society of America* **89**: 395-411.

Flege, J., & Bohn, O-S. (1989). *The perception of English vowels by native speakers of Spanish*. Paper presented at the meeting of the Acoustical Society of America, Baltimore.

Flege, J., & Eefting, W. (1986). Linguistic and developmental effects on the production and perception of stop consonants. *Phonetica* **43**: 155-171.

Flege, J., & Eefting, W. (1987). Production and perception of English stops by native Spanish speakers. *Journal of Phonetics* **15**: 67-83.

Flege, J., & Hillenbrand, J. (1984). Limits on phonetic accuracy in foreign language speech production. *Journal of the Acoustical Society of America* **76**: 708-721.

Flege, J., & Port, R. (1981). Cross-language phonetic interference: Arabic to English. *Language and Speech* **24**: 125-146.

Garnes, S. (1977). Some effects of bilingualism on perception. *OSU Working Papers in Linguistics*. **22**: 1-10.

Gottfried, T., & Beddor, P. (1988). Perception of temporal and spectral information in French vowels. *Language and Speech* **31**: 57-75.

Jamieson, D., & Morosan, D. (1986). Training non-native speech contrasts in adults: acquisition of the English /θ/–/ð/ contrast by francophones. *Perception & Psychophysics* **40**: 205-215.

Krashen, S. (1973). Lateralization, language learning, and the critical period: Some new evidence. *Language Learning* **23**:63-74.

Kuhl, P. (1987). Perception of speech and sound in early infancy. In P. Salapatek & L. Cohen (Eds.), *Handbook of infant perception, Vol. 2*. Pp. 275–382. Orlando: Academic Press.

Lado, R. (1957). *Linguistics across cultures*. Ann Arbor: University of Michigan Press.

Lenneberg, E. (1967). *Biological Foundations of Language*. New York: Wiley.

Logan, J., Lively, S., & Pisoni, D. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America* **89**: 874-886.

MacKain, K., Best, C., & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics* **2**: 369-390.

McClaskey, C., Pisoni, D., & Carrel, T. (1983). Transfer of training of a new linguistic contrast in voicing. *Perception & Psychophysics* **34**: 323-330.

Mitleb, F. (1981). Timing of English vowels spoken with an Arabic accent. *Research in Phonetics*, Report No. 2, Department of Phonetics, Indiana University, Bloomington.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A., Jenkins, J., & Fujimura, 0. (1975). An effect of linguistic experience: the discrimination of /r/ and /l/ by native speakers of Japanese and English. *Perception & Psychophysics* **18**: 331-340.

Morosan, D., & Jamieson, D. (1989). Evaluation of a technique for training new speech contrasts: generalization across voices, but not word position or task. *Journal of Speech and Hearing Research* **32**: 501-511.

Obler, L. (1982). The parsimonious bilingual. In L. Obler, & L. Menn (Eds.), *Exceptional Language and Linguistics* (pp. 339-346). New York:Academic Press.

Olson, L., & Samuels, S. (1982). The relationship between age and accuracy of foreign language pronunciation. In S. Krashen, R. Scarcella, & M. Long (Eds.), *Child–adult differences in second language acquisition.* Rowley Mass.: Newbury House.

Oyama, S. (1982). A sensitive period for the acquisition of a nonnative phonological system. In S. Krashen, R. Scarcella, & M. Long (Eds.), *Child–adult differences in second language acquisition.* Rowley Mass.: Newbury House.

Pisoni, D., Aslin, R., Perey, A., & Hennesy, B. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance* **8**: 297-314.

Port, R., & Mitleb, F. (1983). Segmental features and implementation in acquisition of English by Arabic speakers. *Journal of Phonetics* **11**: 219-229.

Purcell, E. & Suter, R. (1980). Predictors of pronunciation accuracy: a reexamination. *Language Learning* **30**: 271-287.

Scovel, T. (1988). *A time to speak: a psycholinguistic inquiry into the critical period for human speech*. New York: Harper & Row.

Sinnott, J. (1989). Detection and discrimination of synthetic English vowels by old world monkeys (Cercopithecus, Macaca) and humans. *Journal of the Acoustical Society of America* **86**: 557-565.

Snow, C. (1987). Relevance of the notion of a critical period to language acquisition. In M. H. Bornstien (Ed.), *Sensitive periods in development: interdisciplinary perspectives* . Hillsdale, NJ: Lawrence Erlbaum Associates.

Strange, W., & Broen, P. (1980). Perception and production of approximant consonants by 3-year-olds: a first study. In G. Yeni-Komshian, J. Kavanagh, & C. Ferguson (Eds.), *Child phonology, Vol. 2: Perception.* New York: Academic Press. Pp. 117–154.

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception and Psychophysics* **36**: 131-145.

Suter, R. (1976). Predictors of pronunciation accuracy in second language learning. *Language Learning* **26**: 233-253.

Tahta, S., Wood, M., & Loewenthal, K. (1981). Foreign accents: factors relating to transfer of accent from the first language to a second language. *Language and Speech* **24**: 265-272.

Tees, R., & Werker, J. (1984). Perceptual flexibility: maintenance or recovery of the ability to discriminate non–native speech sounds. *Canadian Journal of Psychology* **38**: 579-590.

Underbakke, M., Polka, L., Gottfried, T., & Strange, W. (1988). Trading relations in the perception of /r/– /l/ by Japanese learners of English. *Journal of the Acoustical Society of America* **84**: 90-100.

Werker, J. (1983). Developmental changes across childhood in the perception of non–native speech sounds. *Canadian Journal of Psychology* **37**: 278-286.

Werker, J. (1986). The effect of multilingualism on phonetic perceptual flexibility. *Applied Psycholinguistics* **7**: 141-156.

Werker, J., Gilbert, J., Humphrey, K., & Tees, R. (1981). Developmental aspects of cross-language speech perception. *Child Development* **52**: 349-355.

Werker, J., & Lalonde, C. (1988). Cross–language speech perception: initial capabilities and developmental change. *Developmental Psychology* **24**: 672-683.

Werker, J., & Logan, J. (1985). Cross-language evidence for three factors in speech perception. *Perception and Psychophysics* **37**: 35-44.

Werker, J., & Tees, R. (1984a). Phonemic and phonetic factors in adult cross–language speech perception. *Journal of the Acoustical Society of America* **75**: 1866-1878.

Werker, J., & Tees, R. (1984b). Cross–language speech perception:evidence for perceptual reorganization during the first year of life. *Infant Behaviour and Development* **7**: 49-63.

Williams, L. (1979). The modification of speech perception and production in second–language learning. *Perception & Psychophysics* **26**: 95-104.

Williams, L. (1980). Phonetic variation as a function of second-language learning. In G. Yeni-Komshian, J. Kavanagh, & C. Ferguson (Eds.) *Child phonology, Vol. 2: Perception* . New York: Academic Press. Pp. 185-215.

# CHAPTER 2
# SPECTRAL AND TEMPORAL CUES IN VOWEL PERCEPTION BY NATIVE SPEAKERS OF ENGLISH AND ARABIC[1]

## Introduction

### *'Quantity' vs 'non-quantity' languages*

The purpose of the study presented in this chapter is to explore the role of first-language experience in the perception of spectral and temporal properties of vowels from a second language. A number of previous studies have addressed this issue, but relatively little work has directly investigated the difficulties faced by speakers of a "quantity language" who learn a second language in which vowel duration differences are of less importance.

A number of languages, often referred to as quantity languages, are known to distinguish certain vowel pairs on the basis of duration, as in Cairene Arabic /si:d/ 'master' vs. /sid/ 'close up.' In measurements of Jordanian Arabic words produced in a sentence context, for example, Mitleb (1981) found the ratio of long to short vowel durations to be on the order of 1.5, and Norlin (1981) found a ratio of approximately 2.0 in data from Egyptian speakers. It would be incorrect to state that the only difference in long-short pairs in quantity languages is duration; in at least some pairs there may also be differences in quality, which may be perceptually relevant. Abramson & Ren (1990), for instance, found that long vowels in Thai were about 1.9 times as long as their short counterparts, but there were also differences in the frequencies of F1, F2, and F3 in 9 pairs.

The spectral properties of Arabic long and short vowels vary from one dialect to another. Norlin (1984) measured temporal and spectral properties of Cairene Egyptian long-short vowel pairs produced in a dental CVC context within a sentence frame.

---

[1] Portions of this research were presented at the 120th meeting of the Acoustical Society of America in San Diego, November 1990.

According to his measurements, the /i:/ and/i/ tokens differed considerably in terms of F1 and F2 frequencies, as did the /u/ and /u:/ tokens, while the /a:/–/a/ pair appeared to occupy roughly the same region in the F1-F2 space. He also reported a long-short duration ratio of roughly 2:1. On the basis of his data it might be hypothesized that Cairene speakers have robust spectral and temporal cues available to them for the high vowels, but that spectral cues are likely to be less useful in distinguishing the /a:/–/a/ pair. Al Ani (1970), who measured productions from the Iraqi dialect, reported essentially opposite results. He determined the F1 and F2 values of vowels uttered in isolation, observing little spectral difference in the /i/–/i:/ and /u/–/u:/ pairs, but a relatively large difference between /a/ and /a:/. He also reported a long-short duration ratio of 2:1. His findings suggest that duration cues are more robust than spectral cues in the dialect considered, and that listeners must therefore rely more extensively on duration cues than on spectral cues in the perception of long-short pairs.

Such measurement data do not, of course, indicate the ways spectral and temporal cues are actually used by listeners in the perception of vowels in a quantity language, although it has been proposed that when duration differences between vowels are reliably present they are likely to be perceptually important, particularly if differences in quality are small (Bennett, 1968; Bohn & Flege, 1990a, 1990b). Abramson and Ren (1990) had 50 native speakers of Thai identify modified natural tokens of 9 short-long vowel pairs from Thai. The short vowels were lengthened by reiterating pitch periods, and the long vowels were shortened by removing them. While the subjects generally showed complete crossovers in identifications as a function of duration, their category boundaries varied, depending on whether the original unmodified stimulus was the long or short member of the pair. The authors concluded that although duration

was the major cue used by Thai listeners in distinguishing such pairs, spectral properties seemed to operate as a secondary cue.

Johansson (1984) explored the role of duration in the perception of long-short pairs of synthetic Swedish vowels and obtained large differences in identifications by native speakers as a function of duration. His conclusion that duration is the major cue in Swedish is plausible but perhaps premature, given that he did not vary spectrum at all in his stimuli. His argument is further undermined by the fact (observed by him) that long vowels in Swedish are typically diphthongs while short vowels are generally monophthongs, although all his stimuli were monophthongal.

While "non-quantity" languages do not have minimal pairs of words distinguished exclusively by vowel duration, they often have other sorts of vowel duration phenomena which may have perceptual significance. English vowels are characterized by both intrinsic and extrinsic differences in duration. 'Lax' vowels are generally shorter than their 'tense' counterparts and low or open vowels have been found to be longer than high or close vowels (Peterson & Lehiste, 1960; Umeda, 1975; Crystal & House, 1988). It is generally believed that these intrinsic differences in duration play a role in vowel identification. For instance, Ainsworth (1972) observed that identification of synthetic English vowels varied according to perceived duration.

In addition to intrinsic differences, the duration of English vowels is known to vary as a function of the voicing of the following consonant, with the ratio of vowel duration before a voiced consonant to that before a voiceless one being on the order of 1.5 (Peterson & Lehiste, 1960; Chen, 1970). Because many other languages are known to show a similar voicing-conditioned vowel duration difference, Chen (1970) has argued that vowel lengthening in the context of a voiced consonant is universal. However, the fact that there are large cross-language differences in the size of the ratio

suggests that production of such differences is at least partially learned. In fact, in studies of Arabic by Mitleb (1981) and Flege & Port (1981) no statistically significant consonant-conditioned differences in duration emerged.

Experiments with synthetic stimuli have revealed that vowel duration can serve as a perceptual cue to voicing in English final consonants (Raphael, 1972). Furthermore, a trade-off has been observed between the effects of vowel duration and consonant closure duration on the perception of voicing. In general, a final synthetic consonant will be perceived as more voiced as the ratio of final consonant duration to vowel duration decreases (Denes, 1955). However, experiments with synthetic speech may not give a true picture of the importance of vowel duration in the perception of natural utterances. When Hogan and Rozsypal (1980) modified the vowel durations of natural tokens, they found evidence that, in at least some cases, other cues to voicing (such as the presence or absence of a voice bar, or the duration of the closure interval) may be more important. For some vowels, such as intrinsically long and high vowels, a change in the perceived voicing characteristic of the following consonant was easier to obtain through vowel duration manipulations than for others. They argued that vowel duration may not be a sufficient cue for voicing in the case of intrinsically short vowels such as /ɪ/ and /ʌ/. Another fact which argues against vowel duration as a robust perceptual cue to consonant voicing in natural speech is that vowel duration differences tend to be very small in non-phrase-final syllables (Umeda, 1975; Klatt, 1976).

### *Spectral and temporal properties of vowels and second language learning*

Numerous studies have shown that second-language (L2) learners' perceptions of speech sounds in their second language are influenced by L1 experience. It seems reasonable to expect that the

ways in which spectral and temporal properties of vowels are perceived might be subject to first language influences. Gottfried and Beddor (1988) had French- and English- speaking subjects identify a synthetic continuum ranging from French *côte* (/o/) to *cotte* (/ɔ/) in which vowel quality (3 steps) and duration (10 steps) were varied orthogonally. They observed that although the native English speakers were strongly influenced by spectral quality, their data showed a significant effect of vowel duration, while the data from the native French speakers showed no such effect. This was so, even though a subset of the native French group *produced* a duration difference between these two vowels. Gottfried and Beddor argued for an overall greater perceptual use of vowel duration on the part of English speakers compared with French speakers because, according to their analysis, the French sound system makes less use of vowel duration differences than the English system. They observed that French has only one common vowel pair in which a duration difference occurs (together with a spectral contrast), whereas English has several 'tense-lax' pairs in which duration is of some importance. They proposed that English speakers are therefore more likely than French speakers to integrate spectral and temporal cues in vowel identification.

Bohn & Flege (1990a) had native speakers of German identify tokens from an English *beat-bit* continuum similar to the stimuli to be used in the present study. Some of the subjects showed heavy use of spectral cues and less use of vowel duration, as did a group of native English speakers. However, some subjects used temporal properties much more than spectral ones. The two patterns observed in the data were not related to the amount of experience the subjects had had with English or the regional dialect of German which they spoke. Rather, it appeared that there were unpredictable individual differences in the degree to which the subjects were sensitive to

duration, which may have corresponded to differences in the perception of the /i/-/ɪ/ difference in their native language.

In another relevant study, Van Heuven (1986) compared labelling patterns on synthetic Dutch vowels differing in quality and duration heard by native Dutch speakers and Turkish speakers of Dutch. His stimuli consisted of bVt and fVt words with 34 spectral values and 6 duration steps. From the data he reports, it appears that the Turkish speakers generally relied more on duration in their judgments than did the Dutch speakers, although Dutch in fact uses duration contrastively. This was especially true in the case of the Dutch /a/ - /a:/ contrast, for which the Dutch speakers made approximately equal use of duration and spectrum, while the Turkish speakers seemed to ignore spectral differences entirely. Van Heuven reports that Turkish does not use vowel duration contrastively, and proposes that there may be allophonic vowel duration differences in Turkish which might explain his findings. Nevertheless, grammars of Turkish (e.g. Underhill, 1975) state that Turkish does indeed have minimal pairs of words distinguished by vowel duration. The situation is further complicated by Van Heuven's finding that, in their productions, the Turkish speakers failed to *produce* as large a duration difference in several vowel pairs as did the native Dutch group. It is difficult to draw any firm conclusions from this study. One problem is the small sample size - only five Turkish listeners participated. Also, no information is given on how experienced the subjects were with Dutch.

Wiik (1965) describes a detailed study of the differences between the sound systems of Finnish (which uses vowel length contrastively) and English in an effort to account for the pronunciation difficulties faced by Finnish-speaking learners of English. As in most contrastive studies of the time, the chief source of information about the types of production errors made by L2 learners is anecdotal information. However, he does report a perceptual experiment

involving about 30 Finnish school children who spoke no English. When asked to identify natural English productions of English vowels as long, neutral, or short, they rated/i/ as long much more often than /ɪ/.

Bennett (1968) argued that linguistic experience was responsible for differences in perception of a novel vowel contrast by 20 native English and 20 native German speakers. The subjects were tested in an ABX discrimination paradigm on synthetic tokens of two back unrounded vowels, one long and one short, in a /sVš/ context. Four spectral and four duration steps were used. In each trial, after hearing the tokens from the extreme ends of the continuum (in terms of both duration and spectrum) the subjects were presented with a third token to be identified as more like A or more like B. Unfortunately, no statistical analysis was carried out, but it appears that the English subjects matched the X tokens slightly more often on the basis of duration than on spectrum, while the German speakers showed a strong tendency to use spectrum more than duration. Bennett also argued, on the basis of an additional experiment that both groups of speakers tended to show more sensitivity to duration in their native language when quality differences were smaller.

Finally, Jonasson and McAllister (1972) had one native English speaker of Swedish produce 17 Swedish vowels in nonsense words within a sentence frame. In general, the subject performed poorly in distinguishing long vowels from short ones. For instance, his productions of /i/ were actually longer than /i:/, even though a native-speaking subject produced a consistent long-short contrast here. Overall, the English speaker's values showed much higher standard deviations over five replications than did the values for a native speaker. These results might be taken to mean that the English speaker had not learned the importance of vowel duration in Swedish, presumably because English is a non-quantity language.

However, the use of only one English speaker in this experiment makes any firm conclusions impossible.

## Experiment I: Identification of a synthetic English /bɪt/-/bɪt/ continuum

In the experiment reported here, a group of native Arabic speakers who had learned English in adulthood and a group of native speakers of English categorized synthetic tokens of English words as either *beat* or *bit*. Both the spectral and temporal characteristics of the vowels were varied orthogonally. The question to be addressed here was whether speakers of a quantity language, in this case Arabic, would show greater sensitivity to temporal cues than native speakers of English when labelling vowels on a *beat-bit* continuum.

### Methods

#### *Subjects*

Participants in this experiment were 23 native speakers of American English and 23 native speakers of Arabic who were recruited mainly from the student and staff populations at the University of Alabama in Birmingham. The native English group (referred to below as the EN group) consisted of 12 females and 11 males between 19 and 51 years of age, most of whom had grown up in the south-eastern United States. The native Arabic speakers (the AR group) had all learned English after the age of 15 and had varying degrees of experience with it. This group consisted of 21 males and 2 females between the ages of 19 and 57 years. Detailed descriptive information about these subjects is provided in Table 2-1. They had come to the United States from a number of Arabic-speaking areas - Kuwait (6), Jordan (5), Sudan (4), Saudi Arabia (3), Syria (2), Palestine (2), and Egypt (1), and had been living in the United States for 1-27 years. Table 2-1 also gives the number of years the subjects

had used English as a language of communication (YOE), their length of residence (LOR) in the United States or other predominantly English-speaking area, the age at which they first began to use English for communication (AOL), a self-estimate of the percentage of time they used English in their daily affairs (%USE), a self-estimate of the accuracy of their English pronunciation on a scale of 1 to 7 (SR; 1=poor, 7=native-like), and a similar estimate made by the experimenter on the basis of a short interview (ER). Since most subjects from both groups were from the university community, the variety of English spoken by the native speakers was fairly representative of the speech the Arabic speakers were likely to be exposed to on a day-to-day basis. The subjects were paid $10 for a one-hour session, which involved a number of short tasks, including the present experiment.

***Stimuli***

An English beat-bit continuum was synthesized using the parallel mode of a Klatt (1980) synthesizer implemented on a Digital PDP-11 computer. The sampling rate was 10 kHz, and parameters were updated every 5 ms. Each stimulus consisted of a 5 ms initial release burst; a vowel portion containing initial transitions appropriate for a /b/, a steady-state interval, and final transitions for /t/; a 65 ms silent period corresponding to the /t/ closure; and a 5 ms burst with 30 ms of aspiration corresponding to the /t/ release.

Six spectral and six temporal steps were synthesized independently to produce a matrix of 36 stimuli whose properties are summarized in terms of relevant Klatt parameters in Table 2-2. The nominal steady state values of F1, F2, and F3 for the six tokens at the end of the continuum corresponding to /i/ were 233, 2400, and 3080 Hz respectively. At the /ɪ/ end of the continuum, these values were 361, 2000, and 2760 Hz. All intermediate values were calculated by linear interpolation. F4 and F5 were fixed at 3800 and 4500 Hz

respectively, and FØ was set to 120 Hz for the first half of each token and then fell linearly over the second half to 100 Hz. The six longest vowels (including transitions) were 250 ms. Each additional step was determined by decrementing the value of the previous step by 25 ms. This interval size was based on Klatt's (1976) estimate of one JND for segmental duration. The six shortest stimuli were therefore 125 ms, giving a 2:1 ratio of the duration of the longest to the shortest stimuli. Formant frequencies were confirmed with LPC analysis and stimulus durations were measured on a CRT display to ensure that the actual stimulus properties were close to the nominal values (see Appendix). All stimuli were normalized for peak intensity before presentation to ensure that they were of about equal loudness.

During the stimulus preparation phase, three native English listeners were asked to perform open-set word identification tasks at different stages in the development of the stimulus set. They listened to each synthetic stimulus once and wrote each word in standard orthography on a sheet of paper. On the basis of their responses, modifications to spectral values were made. The final stimulus set consisted only of words which had been identified as containing either /i/ or /ɪ/ 100% of the time. Errors in the identification of the final consonant averaged less than 5%.

***Procedure***

Listening sessions were conducted in a sound-treated room in the speech laboratory at the Department of Biocommunication, University of Alabama in Birmingham. Stimuli were low-pass filtered at 4.8 kHz with a Krohn-Hite 3202R filter, amplified with a Crown D-75 amplifier, and presented binaurally through Sennheiser HD530 headphones at a comfortable listening level. The subjects, who all passed a hearing screen (500-4000 Hz at 25 dB), were seated in front of a response box with buttons labelled *beat* and *bit*, and were instructed to press the button corresponding to the keyword most like the stimulus. The experimenter confirmed that the native Arabic

speakers were familiar with these words and provided sample sentences containing each one. Data were collected from individual subjects in two blocks, each lasting about eight minutes. In each block, the subjects identified five randomized replications of each of the 36 stimuli for a total of 180 identifications per block. A total of 16560 responses were collected (360 responses per subject x 46 subjects).

## Results

### *Preliminary Analysis*

A visual summary of the data is provided in Figs. 2-1 and 2-2. In Figure 2-1, the percentage of *beat* identifications pooled over the six temporal possibilities is given for each group. As expected, the native English subjects (EN group) showed an almost complete shift from *beat* to *bit*, with *beat* identifications ranging from 97.0% at the first step to 2.8% at the last step. Scores from the Arabic speakers (AR group) ranged from 83.3% to 23.0% and therefore illustrate a slightly less complete shift from one category to the other. Figure 2-2 shows the scores pooled over spectral steps. Values from the EN group ranged from 38.6% at step 1 (the shortest stimulus) to 51.5% at step 6 (the longest one). The AR group showed a wider range of scores, from 17.0% to 75.7%.

An impressionistic examination of these data suggests that the Arabic speakers relied more on the durational properties of the stimuli in making their category assignments than did the native English speakers. The steepness of the function representing the English speakers' spectral data and the more horizontal shape of the function for duration steps suggests that the EN group were very sensitive to spectral differences in the stimuli, but made less use of durational information. The Arabic speakers, showed approximately the same amount of shift along the spectral dimension as along the temporal one.

An examination of the change in *beat-ID* scores as a function of stimulus duration at each spectral step (Table 2-3) shows that, in terms of absolute %-shift, the largest effect of duration was at spectral step 3 in both groups: from 17% *beat ID* on the shortest stimulus to 97% (i.e., an 80% shift) on the longest in the AR group; and from 47% to 78% (a 31% shift) in the EN group. In both groups then, fairly large effects of duration are seen in the case of a stimulus item which is fairly ambiguous spectrally. However, as can be seen from Figure 2-3, at any given spectral step, the EN group showed much less change as a result of differences in vowel duration than did the AR group. At any given spectral step, relatively little change is seen over duration steps in the EN listeners' function, whereas a noticeable change is seen between steps 1 and 3 in the AR listeners' data, regardless of spectral step. The fact that the functions in each of the panels of Figure 2-3 are fairly parallel suggests that, within groups, the effects of duration are relatively constant across spectral steps.

***Linear Logistic Analysis***

A more detailed analysis of the data was obtained by submitting them to a linear logistic analysis (Nearey, 1990). In this procedure equations are generated for the logs of the probability ratios (logit scores) of the categories in question. It is expected that the entire response surface for each subject's data can be analyzed adequately by assuming an independent contribution from each variable to the evaluation function. Three models represented by the following linear evaluation functions were considered:

$$f(r,s) = b(r) + a_1(r)\,DUR + a_2(r)\,SPEC \qquad (1)$$

$$f(r,s) = b(r) + a_1(r)\,DUR \qquad (2)$$

$$f(r,s) = b(r) + a_1(r)\,SPEC \qquad (3)$$

where $b\,(r)$ represents an overall bias term and $a_1\,(r)$ and $a_2\,(r)$ are the coefficients of stimulus-tuned terms corresponding to temporal and spectral steps respectively. Model 1 above is a prediction equation which takes into account both the temporal and spectral properties of the stimuli, while the other two models include only one of the two possible stimulus-tuned terms. Prediction equations were generated for individual subjects, and a pooled analysis was carried out for the two groups. It was then possible to compare Model (1) with Models (2) and (3) for each of the two groups by computing an $F$ - ratio of the G2 deviance statistic (likelihood ratio Chi-square) to the overdispersion factor (Nearey, 1990). Table 2-4, summarizes the results of this analysis. For the AR subjects Model 1 proved to be significantly better than either Model 2 ($F$ (1,33)=148.42, $p < 0.001$) or Model 3 ($F$ (1,33)=125.74, $p < 0.001$). For the EN subjects, again Model 1 was better than Model 2 ($F$ (1,33)=808.87, $p < 0.001$) or Model 3 ($F$ (1,33)=23.85, $p < 0.001$). In short, a model which takes into account both the durational and the spectral properties of the stimuli gives the best account of the results for both groups of subjects.

By substituting the appropriate $b$ , $a_1$ , and $a_2$ values for each group into equation 1, it is possible to represent predicted responses in territorial maps reflecting the trading relationship between duration and spectrum in the subjects' responses (see Figure 2-4). When the function is equated to 0, it gives a line along which the probabilities of the *beat* and *bit* categories are equal. If the equation is solved for SPEC as follows:

$$SPEC = -(b(r) + a_1\,(r)DUR) \,/\, a_2\,(r) \qquad (4)$$

It can be seen that the slope of the line specified is determined by the expression

$$-a_1\,(r) \,/\, a_2\,(r) \qquad (5)$$

Fig. 2-4(a), representing the EN group, shows a relatively horizontal function, while Fig. 2-4(b) shows a function rising to the right. The regions on each map are the areas in which one vowel category predominates. From Fig. 2-4(a) it can be seen that for the EN group the relative effect of spectrum is much greater than the effect of duration. In the case of the AR group, the importance of the two types of cues is about the same.

To explore further the relative importance of the two cues in the data from each group, ratios of the two coefficients from individual subjects' functions were calculated according to (5) above. These ratios, which may be regarded as a measure of the relative sensitivity of the subjects to the two types of cues, are given in Table 2-1, and will be referred to as the "relative sensitivity measure." The mean value for the EN group was 0.007 (s.d. = 0.005); for the AR group it was 0.101 (s.d. = 0.121). The two groups were then compared using a two-sample *t*-test, which revealed that the AR data in general showed larger ratios computed according to expression (5) than the EN data ($t(44)=3.746$, $p< 0.001$; two-tailed). In other words, the AR functions did indeed show a greater relative contribution of duration than did the EN functions.

***Within-Group Variation***

The error bars in Figure 2-1, which represent standard deviations, suggest that there is generally more variability in the Arabic data than in the English data. Data pooled over temporal steps show the greatest variability in the AR values at steps 4, 5, and 6, where the mean standard deviation is 13.2%. In the EN group the greatest variability (a mean of 8.3) is seen in the spectrally ambiguous tokens at steps 3 and 4. An examination of the standard deviations in the data pooled over spectral steps reveals fairly constant within-group variability at all steps for both the AR group

(with a mean of 9.4) and the EN group (4.4), with the AR group showing standard deviations more than twice as high as those of the EN group.

An examination of the *Relative Sensitivity (RS)* column of Table 2-1, indicates considerable variation in the data from the individual AR subjects. The ratios range from very "native English-like" values of about 0.004 to much larger numbers which reflect heavy use of duration. The individual differences are illustrated in two ways in Figures 2-5 and 2-6. Figure 2-5 summarizes the data from individual subjects in terms of the %-change in *beat* identifications from step 1 to step 6 on both the spectral (pooled over temporal) and temporal (pooled over spectral) dimensions. Data points in the upper left hand corner indicate a greater change due to spectral cues relative to temporal ones. Naturally, most of the EN scores are clustered in this region. However, seven of the AR scores are located to the left of the data point from the EN subject who showed the largest shift due to duration (40%)[2]. These subjects are the seven subjects at the top of Table 2-1. They categorized the stimuli in much the same way that the EN subjects did.

The remainder of the AR subjects showed varying degrees of use of spectrum and duration. Several additional subjects came close to the EN pattern, but data from others appear in the lower right hand corner; the latter apparently relied almost exclusively on duration.

In Figure 2-6, the coefficients from the linear logistic analysis for individual subjects have been plotted. The coefficients have been normalized by dividing by the largest value observed on each dimension, so that the largest value on both dimensions is 1. It can be seen that most of the data points from the EN subjects are clustered toward the left-hand side of the figure, indicating greater

---

[2] One of these points, located at (12,97) is obscured on the figure by other points near the same location.

effects of spectrum than duration and greater effects of spectrum than in the AR group. It can also be seen that while data points from some of the AR subjects appear in the region occupied by the data points from the EN group, most points fall outside this area. At the bottom of the figure are data from several AR subjects who showed relatively small effects of spectrum, but exhibited great variation in relative sensitivity to duration.

It is not obvious why there is so much variability in the data from the AR group. It might be hypothesized that dialectal factors are playing a role here, but when the ten subjects whose data show the most English-like pattern were identified from Table 2-1, three were found to be from Kuwait, two were from Saudi Arabia, two were from Jordan, and one each was from Syria, Palestine, and Egypt. In contrast, of the ten subjects showing the least English-like pattern (bottom of Table 2-1), four were from the Sudan, three were from Jordan, and one each was from Saudi Arabia, Kuwait, and Syria. The only hint of dialectal differences here was that all four subjects from the Sudan showed strong duration effects.

Another possibility is that amount of experience with English determined the degree to which the Arabic speakers used duration cues in their categorizations. There is no evidence, however, that this is the case. An examination of Table 2-1 makes this quite clear: several of the subjects near the bottom of the table (i.e. those whose data differed most from the native English data) have more than ten years of experience with English. A Pearson correlation coefficient of only 0.049 (ns) was computed between the variable YOE and the ratios of the duration and spectrum coefficients in the linear logistic analysis.

There is also considerable variability in the data from the EN group. From Figure 2-5 it can be seen that while all subjects but one showed a %-change of 85% or more due to spectral differences, the %-change due to duration ranged from –15% to +40%. Also Figure 2-6

shows data points which cover a large vertical (but smaller horizontal) area. This indicates that relative sensitivity to the spectral cues varies considerably even within the EN group. Since the native English subjects were a relatively homogeneous group, it seems reasonable to conclude that the within-group differences seen here are at least partly the result of individual differences in vowel perception.

**Discussion**

In this experiment, native speakers of English and Arabic showed differences in sensitivity to spectral and temporal cues when labelling tokens from a synthetic English *beat-bit* continuum. For the native English speakers, whose results were strongly influenced by the spectral properties of the stimuli, the duration cue was generally much less important than for the Arabic speakers, most of whom relied greatly on duration and less than the native English speakers on spectrum in making their identifications. These results indicate effects of first language experience on vowel perception. In general it does appear that knowledge of a quantity language may result in greater sensitivity to duration differences in the labelling of at least this type of continuum. Nonetheless, the precise mechanisms which led to the differences seen here are unknown.

A key question which must be asked here is whether the Arabic L2 learners had actually established distinct representations for the English vowels /i/ and /ɪ/ . That is, did they recognize that English /i/ and /ɪ/ are distinct from similar vowels in Arabic? Some researchers have tried to answer this type of question by attempting to induce "linguistic set" effects in L2 learners. Caramazza, Yeni-Komshian, Zurif, & Carbone (1973), for instance, were unable to obtain such an effect when they had bilingual listeners label the same VOT continuum in "English" and "French" conditions. Instead, their listeners appeared to use VOT values

intermediate between French and English norms in their labelling patterns in both conditions. This may have meant that the listeners had not established distinct representations for the consonants from the two languages. Elman, Diehl, & Buchwald (1977), on the other hand, did report a language set effect in data from *some* of their Spanish-English bilinguals who identified natural tokens of stops with varying degrees of VOT. The strength of the effect was related to fluency in English.

In the present experiment, most of the non-native subjects failed to identify the /i/–/ɪ/ continuum in a manner which could be considered "native-like." Several of them, in fact, exhibited a highly inappropriate use of duration and very little use of spectral properties, which indicates that they had probably not developed accurate representations of English /i/ and /ɪ/. One possible explanation of the performance of these subjects is that they did not make use of an English "mode" of perception at all, but rather that, because of the effects of equivalence classification (Flege, 1987) they simply identified the stimuli as if they were distorted exemplars of the Arabic vowels /i:/ and /i/, which are known to differ much more in duration than the corresponding English categories. In other words, the long stimuli may have sounded to them much like Arabic /i:/, while the short stimuli may have sounded like Arabic /i/. Also, for most subjects, the stimuli near the two spectral extremes may have sounded like exemplars of the two Arabic vowel categories. The Arabic subjects' categorizations, then, may illustrate a "matching" strategy, whereby the subjects grouped the stimuli by matching them to one of two similar Arabic vowel categories. This would explain their strong reliance on duration. This account will be referred to below as the "equivalence classification proposal."

On the other hand, the performance of a few of the AR subjects was indistinguishable from that of the native speakers. These subjects may indeed have learned enough about the temporal and

spectral properties of the English /i/-/ɪ/ distinction to be aware that these vowels are different from the analogous Arabic vowel categories.

A second proposal which might explain the results presented here is that they reflect global differences in the perceptual strategies used by the two groups. According to this account, knowledge of a vowel system which makes extensive use of duration contrasts may predispose speakers of Arabic to be generally more sensitive to duration cues in vowel perception than are English speakers and speakers of other non-quantity languages. As a result, they might be expected to assign a great deal of importance to duration cues when such cues are available. The results of Experiment I, then, might reflect such global differences in attention. This explanation will be referred to here as the "featural sensitivity" proposal. It is compatible with the view that speakers of Arabic may be aware of a binary opposition between long and short vowels in their native language and that this knowledge is accessed in their perceptions of L2 vowel pairs which happen to differ in duration.

Such an account is similar to the one offered by Gottfried and Beddor (1988), who had native English and native French subjects identify tokens from a French /o/-/ɔ/ continuum. The data from their English listeners showed clear effects of duration, while the French subjects' data did not. This difference may indicate that English speakers generally attend more to temporal properties of vowels than do French speakers, because properties of the English vowel system predispose them to do so. However, this is not the only possible explanation for their results. In fact, the equivalence classification proposal presented earlier for Experiment I might well be appropriate here.

In particular, English speakers might associate the French vowels /o/ and /ɔ/ with a vowel pair from English with which they are already familiar. One potential candidate is the English /o/-/ʌ/

pair, which is distinguished by a noticeable duration difference, as well as a spectral difference. If the English listeners associated /o/ and /ɔ/ with these vowels, then an inappropriate use of duration cues might be observed in an identification experiment.

Qualitatively speaking, it appears, then, that the results of both Experiment I and those of Gottfried and Beddor (1988) can be accounted for by either of the proposals discussed above. It is also conceivable that the listeners in these studies were influenced *both* by a general preference for duration cues and by equivalence classification, or that neither of these explanations is correct. However, the two accounts proposed do make slightly different predictions about the ways speakers of Arabic might perceive vowels from English or other languages. If the equivalence classification proposal is correct, then we might expect speakers of Arabic to exhibit greater use of duration only when they are able to "match" a particular vowel distinction to a long-short distinction in their native language. This might not occur, for instance, if two L2 vowels exhibit single-category assimilation as described by Best (1990), and are therefore *both* heard as the same L1 vowel. In contrast, the featural sensitivity proposal predicts that speakers of Arabic should *always* be observed to pay more attention to vowel duration differences which cue some distinction than should speakers of English. For instance, they should readily notice the effects of consonant voicing on the duration of a preceding vowel in English, even though this effect is non-significant in Arabic. While there are no data available on Arabic speakers' perceptions of voicing-conditioned vowel duration differences, it has been shown that Jordanian and Saudi speakers do not *produce* as large a duration difference between English vowels before voiced and voiceless consonants as do English speakers (Mitleb, 1981; Flege & Port, 1981). This finding may indicate a lack of awareness of such duration differences and may therefore discredit the featural sensitivity

account. However, such evidence is indirect and must be viewed with caution until perceptual data are obtained.

*Individual Differences*

Notable individual differences were observed in the results of this experiment. Although the Arabic speakers generally tended to make greater use of temporal information, several of them performed the task in much the same way that the native English speakers did; in other words, they relied more heavily on spectral cues than on temporal cues. In fact, seven subjects, whose data appear in the upper left-hand corner of Figure 2-5 did not appear to differ from the native speakers in their performance on this task. An examination of the subjects' linguistic backgrounds did not yield any strong evidence that the individual differences were related to dialectal differences in the subjects' native language, although all listeners from the Sudan showed similar patterns. Nor did it reveal any correlation between amount of experience with English and use of temporal information.

One possible explanation for the within-group variation seen here is that it reflects individual variation in the learning of a non-native sound system in adulthood. That is, some adult L2 learners acquire their new sound system in a more native-like way than others. It may be the case that the subjects who performed like the native speakers were more successful than the other subjects at learning the English /i/–/ɪ/ distinction. However, a second possibility is that the differences in duration use reflect idiosyncratic perceptual strategies; some listeners may, for unknown reasons, choose to use duration cues more than other listeners when both spectral and duration differences exist between two vowel categories. In fact, Bohn & Flege (1990a) proposed a similar explanation to account for individual differences among German speakers labelling continua with spectral and temporal differences. Further support

for this latter account comes from the fact that the English subjects in the present experiment showed varying degrees of sensitivity to duration as well. Differences among these subjects cannot be due to the effects of second language learning because all but one of the subjects were monolinguals. The remaining subject spoke some Spanish, but Spanish does not use duration differences to cue vowel identity (Bohn & Flege, 1990b).

To summarize, the individual data presented here reflect up to three types of influences. One is an effect of first language: speakers of Arabic used duration more than speakers of English in labelling a *beat-bit* continuum. The second is an effect of success in L2 learning. Some of the Arabic speakers labelled the English continuum in a more native-like way than others. The third is an effect of individual perceptual strategies. Some speakers, regardless of first language, tended to make more use of temporal cues than others.

Experiment I raises some interesting questions about how the speakers of the two languages might differ when they learn new vowel contrasts. In this experiment the native speakers had a strong advantage. They were asked to identify vowels representing two categories with which they had been familiar all their lives. While the non-natives were expected to perform what was nominally the same task, most of them did not appear to have native-like knowledge of the two English vowel categories. This raises the question of how the two groups might have performed if they both had to learn a new vowel contrast from a third language which involved both temporal and spectral differences. In such a situation, the two groups might be on a more equal footing in the sense that neither would have any a priori knowledge of the relative importance of the two types of cues to the vowel contrast.

## Experiment II: Identification of two foreign vowel categories by speakers of Arabic and English

In Experiment II native speakers of English and Arabic were given a problem–solving task in which they were trained to identify synthetic /bVt/ words containing exemplars of an unfamiliar vowel contrast. The vowel categories chosen were the French vowels /y/ and /ø/, which do not occur in either English or Arabic. The two categories were synthesized in such a way that they differed not only in terms of spectral properties, but in duration as well[3]. The question to be explored here was whether the two groups would perform the same way or whether differences attributable to L1 would emerge.

### Methods

#### *Subjects*

The subjects in this experiment were the same 23 native speakers of Arabic and 23 native speakers of English who participated in Experiment 1. None spoke a third language (such as French) with a contrast between the vowels /y/ and /ø/.

#### *Stimuli*

Two sets of 13 /bVt/ stimuli were synthesized using the same equipment and software as in Experiment I. One set had vowel formants appropriate for the French vowel /y/ while the other had formants appropriate for French /ø/. The consonant portions and transitions were synthesized in exactly the same way as in the tokens in Experiment I. It was decided that CVC stimuli were more appropriate than isolated vowels, because the former might be more likely to encourage a truly phonetic level of processing.

---

[3]The difference in spectral properties is characteristic of native French speakers' productions of these vowels, while the difference in duration is not.

Figure 2-7 gives a detailed visual representation of the locations of the synthetic tokens in an F1–F2 space, and Table 2-5 gives the exact formant values. The vowel formant frequencies for the stimuli at the category centres were based on measurements of French vowels by Debrock & Forrez (1976). Before the remaining stimulus values were calculated, the Hz values of the centre stimuli were converted to mels according to the formula

$$f_{mel} = (1000 / log(2)) * log(1+(f_{Hz} / 1000)) ,$$

where $f_{mel}$ = frequency in mels and $f_{Hz}$ = frequency in Hz. All distances were calculated in mels rather than Hz because the mel scale has been argued to reflect perceptual distances more accurately (see e.g., Greiser & Kuhl, 1989). For the /y/ category, the centre stimulus had F1 and F2 values of 297 and 1828 Hz respectively. The /ø/ centre had values of 366 and 1462 Hz. As can be seen from the figure, the category centres were located on a line which also passed through two of the training stimuli from each set. The centres were 237 mels apart with respect to F1 and F2, and the two closest stimuli from the two sets were 117 mels apart. Each set consisted of 5 training stimuli (one at the category centre and 4 variants) and 8 test stimuli (see below). The training stimuli were located at a distance of 60 mels from the centres of their respective categories in four directions, on vectors of 0, 90, 180, and 270 degrees relative to the line passing through the two category centres. The test stimuli were at the same distances but on vectors of 45, 135, 225, and 315 degrees.

For the first half of each stimulus, FØ was set at 120 Hz. It then fell linearly over the second half to 100 Hz. F3 was set to 2137 for the /y/ stimuli and to 2290 for the /ø/ tokens. F4 and F5 values were fixed at 3300 and 3850 Hz for both sets. The /byt/ training tokens had vocalic portions (including transitions) lasting 200 ms while the

vowel duration of the /bøt/ training tokens was 125 ms. All stimuli were normalized for peak intensity before presentation.

As indicated in Figure 2-7, the test tokens had similar spectral properties to the other members of their category. In particular, they were the same distance from their category centres as the training stimuli. These 8 tokens were divided into two sub-categories. Four of them - labelled the "A" tokens - had durations identical to the training tokens of their spectral category. The "B" tokens, however, differed from the training tokens in that they had duration appropriate for the opposite spectral category. The four /byt/ A tokens, for instance, had vocalic portions of 200 ms, while the four B tokens had a vowel duration of 125 ms.

***Procedure***

*Selection of stimuli*

Pilot tests were run before the final stimuli were determined. The training tokens were synthesized at various distances from the centre stimulus until it appeared that the within-category stimuli were discriminable, but not so different that listeners would have difficulty sorting them into two categories. At various stages, tests were run on four native English listeners who knew nothing of the purpose of the experiment to ensure that subjects were able to perform the task. The listeners were simply asked to sort the stimuli into two categories by pressing one of two buttons on a response box. Correct responses were reinforced with a small light which was illuminated when the correct button was pressed. With the final stimulus set, the pilot subjects reached overall correct identification rates of 90% to 100% in a ten-minute training session.

It was also necessary to ensure that the subjects could perceive within-category differences among stimuli in order to encourage subjects to abstract a category rather than simply memorize a single token of each vowel. Accordingly, two native English pilot listeners

were tested on a same-different task in which they classified random pairs of tokens. They were presented with four replications of 16 *same* pairs of tokens and 16 *different* pairs. The latter included all possible pairs of adjacent stimuli (8 pairs, including the test stimuli) from the ring around the category centre and all pairings of the stimuli from on the ring together with the centre stimulus (8 pairs as well). Only the training and A tokens were used, since the B tokens differed noticeably in duration from the other members of their category. It was reasoned that if the subjects could hear the differences between the (spectrally close) tokens in the same–different task, they would also be able to hear the difference between other pairs within each set. The stimuli were presented with feedback in two sets, one for each category. Each of the 32 pairs was presented 4 times, but responses to the first 16 pairs were not counted. Since the two listeners received correct overall discrimination scores of 83% and 78%, it was concluded that the stimuli were sufficiently discriminable from one another.

*Training Phase*

The experiment was carried out in the same acoustically-treated room and with the same equipment as that used in Experiment 1. Subjects were seated in front of a response box with two labelled buttons - one marked with a blue square and the other with an orange square. They were told that they would hear several examples of two foreign words that do not occur in English (or Arabic, in the case of the Arabic-speaking subjects) and that the object of the experiment was for them to determine which words were 'blue words' and which were 'orange words.' They were advised that the difference between the two words was in the vowel and were instructed to begin by listening very carefully to the vowel portion of each token and then to guess which word they had heard by pressing one of the buttons. They were assured that they would soon be able to

tell the difference between the two sets of words. Correct responses were reinforced with a small light which came on immediately after the button was pressed. If the response was correct, a light near the pressed button came on. For an incorrect response, a light near the opposite button was lit. The subjects were forewarned that in the second stage of the experiment the feedback light would be turned off, but that they were to continue responding as in the first part.

The 10 training tokens (five from each category) were presented in blocks of 30 stimuli (3 replications of each stimulus) with a short break of about one minute between blocks. Criterion for moving on to the test phase was set at a minimum of 27/30 correct (90%), with the provision that every subject would listen to at least three training blocks (for a total of 90 trials with feedback) before going on to the test phase. The subjects were monitored on–line to determine when they were ready to go on to the test phase.

*Test Phase*

The test phase, which consisted of a single block of 140 items was administered within two minutes of the last training block. In this phase, the subjects were presented with the ten original training tokens together with the 16 new tokens not used during the test phase. No feedback was provided, and the subjects were not advised that they would be hearing some new tokens along with the previously–heard training stimuli. In order to ensure that the subjects remained confident about their responses during the initial part of this no–feedback phase, the first 10 tokens presented were all simply repetitions of the category centres. It was expected that the subjects would have no difficulty with this set of 'transitional' tokens and would be less likely to become confused when the feedback stopped than they would if they were immediately presented with new stimuli. After the transitional set, the 26 stimuli were each

presented five times for a total of 130 test–phase stimuli. (The responses to the transitional tokens were not used in the analysis.)

**Results**

Of the original set of subjects recruited, three (all from the EN group) failed to reach criterion even after 5 training blocks. These subjects were dropped and replaced by three new subjects. For each remaining subject a *%–error score* was determined by calculating the percentage of times that the training tokens were misidentified in the test phase. At this point it was noted that two subjects - one Arabic and one English - showed error rates of over 30%. This level was deemed unacceptable because it was substantially higher than the rates of most other subjects, and the data from these two subjects were not included in the analysis. The mean error rates for the remaining subjects were 7.2% for the English speakers and 5.7% for the AR speakers, indicating that, for the most part, the subjects continued to perform above the criterion level of 90% on the training stimuli during the test phase. The difference in the error rates for the two groups was not significant ($t$ (43) = 0.723, two-tailed).

Of the 44 participants included in the final analysis, 36 reached criterion within three training blocks (16 English and 20 Arabic) and the 8 remaining subjects reached criterion after four blocks.

The interesting dependent variable here is the number of times the B tokens were identified as belonging to the category agreeing in duration but differing in spectrum. The percentage of times (based on a total of 40) that each subject made such an identification will be referred to here as the subject's 'B' score. A large value for this score suggests that a subject was particularly sensitive to the temporal properties of the synthetic vowels in this experiment. A smaller number indicates that the subject was less sensitive to temporal differences.

Also computed for each subject was an 'A' score, which was the percentage of times the eight A stimuli were mistakenly identified as belonging to the category differing in spectrum and duration. The A condition is, therefore, a baseline condition against which the B scores can be compared. There is no reason to expect this value to differ from the %–error value on the training stimuli since the A stimuli were very similar to the training tokens. In fact there was no significant difference in error rates on the training and the A stimuli ($t$ (43) = 1.186, two-tailed).

Table 2-6 gives the mean 'A' and 'B' scores for the EN and AR groups. For the EN group the B scores range from 10% to 75% with a mean of 44.1% and a standard deviation of 17.3. For the AR group, the range is 35% to 100%, with a mean of 86.9% and a standard deviation of 18.0. A two–way repeated measures analysis of variance (Native Language, Test Condition) was performed with 2 levels of test condition (A and B). The results showed highly significant effects of Language ($F$ (1,42)=64.65, $p < 0.001$) and Test Condition ($F$ (1,42)=317.29, $p < 0.001$), as well as a significant L X T interaction ($F$ (1,42)=49.97, $p < 0.001$). An a posteriori Tukey (a) procedure (Winer, 1971) indicated that the interaction was due to a significant difference between the two language groups in the Test B condition ($p < 0.01$) but not the Test A condition. Both the EN group and the AR group showed a significantly higher B score than A score ($p < 0.01$). This indicates that both groups were influenced by duration to some degree in labelling the B stimuli. However, the AR group showed a higher B score than the EN group. There was no significant difference on the A scores.

Data for individual subjects are provided in Table 2-7. Here the percentage of matches according to duration is given for individual vowel categories as well as the two categories together. For the most part, the EN subjects showed only a moderate tendency to match the test tokens to the category agreeing in duration with many showing

scores of about 40-60%. In contrast, the majority of the AR subjects showed a very strong tendency to do so. Although one subject (AR7) had an overall score of only 35%, a large number of the AR subjects scored between 37 and 40 out of a maximum of 40. In the EN group, there was a slight tendency to match /y/ more than /ø/ with the category agreeing in duration. This was chiefly because five subjects showed very low (10% or less) duration-matching scores on /ø/. This trend proved to be non-significant overall ($t$ (21) =1.658; two-tailed).

During the short debriefing which took place after the task was performed, the subjects were asked to characterize, as best they could, the difference between the "blue" and "orange" words. Only three of the English listeners commented that the blue words were longer than the orange ones. The others usually gave descriptions which referred to vowel quality. For instance, several subjects described the /y/ words as something like *beat* , and the /ø/ words as something like *but* . Almost all the native Arabic speakers (86%), however, stated that one of the words was long and the other short.

***Comparison of results with those of Experiment I***

The Kendall rank correlation coefficient (Ferguson & Takane, 1989) between the AR subjects' relative sensitivity values calculated in Experiment I (see Table 2-1) and the duration–matching scores from Experiment II was computed at 0.43 ($p$<0.05). This value indicates a moderate correlation between the subjects' ranks in the two experiments. It appears, then, that the AR subjects who attended most to duration in Experiment I also did so in Experiment II.

**Discussion**

In Experiment II, the subjects were asked to perform a rather different task from that performed in Experiment I. In the latter,

they were instructed to assign synthetic tokens to familiar vowel categories. In Experiment II, however, they were not told in advance what vowels they were listening for. Instead they were trained to identify long exemplars having one spectral configuration as belonging to one arbitrarily-named category and short exemplars with different spectral properties as belonging to another. They were then presented with new tokens containing conflicting cues and were asked to match these with one of the original categories. Their category assignments on these new tokens, then, reflect their understanding of which of the two dimensions - spectral or temporal - was the more important one distinguishing the two categories.

The results of Experiment II, like those of Experiment I, showed clear differences in the use of spectral and temporal cues between the two groups of listeners. The Arabic group showed a very strong tendency to prefer the temporal dimension in identifying the new tokens. In fact, they seem to have largely ignored the spectral differences in the test stimuli and grouped together spectrally different stimuli which shared the same temporal properties. The native English group, on the other hand, matched the test tokens to the categories much less consistently. Their data appear to have been influenced by both cues. Some of them may have been confused by the test tokens, assigning them sometimes to one category and sometimes to the other, even though they continued to perform correctly on the training stimuli during the test phase. This would indicate that the native English subjects were attending to both the spectral and temporal properties of the stimuli.

However, there is considerable variability in the English listeners' data. It is especially striking that a few of the English subjects seemed not to attend to duration at all, and that during debriefing very few subjects mentioned noticing a duration difference. Yet in an informal test given to three native English listeners who had not participated in the experiment, all could

readily identify (with nearly 100% accuracy) which stimuli were long and which were short when they were told to listen for a duration difference.

Since the two groups of subjects were given identical instructions, the results of Experiment II must be attributable to L1 experience. In attempting to explain the differences, one might propose accounts parallel to those given for Experiment I. One possibility is that both groups of subjects identified the stimuli in this experiment by associating them with two vowel categories in their native language. In the case of the English speakers, differences in both spectrum and duration would presumably have been relevant in the native distinction, and for the Arabic speakers the corresponding native distinction would have involved a duration difference.

As in Experiment I, it is also possible that the greater use of duration by the Arabic speakers reflects a greater overall attentiveness to temporal properties of vowels than might be expected from native English speakers.

These results suggest that speakers from native English and native Arabic backgrounds might use different strategies in acquiring the sound system of another language. In particular, speakers of Arabic might find it easier to exploit duration differences in vowels from a new language and, as a result, be more successful at learning vowel contrasts in another language which are based on duration differences. What cannot be ascertained here, however, is whether such a tendency is due to equivalence classification or to global differences on the part of listeners in use of duration cues.

This study leaves many interesting questions unanswered. In Experiments I and II, only a small portion of the vowel space was considered. Further experimentation with other vowel pairs is necessary in order to gain a full understanding of the differences in vowel perception by speakers of the two languages considered here. Another question which might be raised is whether the Arabic

speakers' greater use of duration in their perceptions of the /i/ - /ɪ/ contrast is related to exaggerations in production. It might be expected that the subjects who show the greatest use of duration perceptually also show exaggerated duration differences in their productions. This issue will be addressed in Chapter III.

## Tables and Figures for Chapter 2

**Table 2-1: AR Subjects in Experiment I (Sorted by Relative Sensitivity to Duration)**

| SJ # | REGION | AGE | YOE | LOR | AOL | %USE | SR | ER | RS |
|---|---|---|---|---|---|---|---|---|---|
| 5 | Kuwait | 27 | 6 | 6 | 21 | 50 | 4 | 5 | 0.0040 |
| 11 | Saudi Ar. | 30 | 11 | 4 | 19 | 100 | 6 | 6 | 0.0084 |
| 7 | Syria | 27 | 7 | 6.5 | 20 | 50 | 4 | 5 | 0.0086 |
| 22 | Saudi Ar. | 19 | 2 | 2 | 17 | 51 | 6 | 6 | 0.0089 |
| 17 | Palestine | 21 | 5 | 3.5 | 16 | 67 | 5 | 5 | 0.0121 |
| 18 | Egypt | 57 | 27 | 27 | 30 | 73 | 6 | 5 | 0.0147 |
| 13 | Kuwait | 24 | 5 | 5 | 19 | 93 | 5 | 5 | 0.0169 |
| 3 | Jordan | 21 | 3 | 2.5 | 18 | 67 | 6 | 4 | 0.0222 |
| 4 | Kuwait | 23 | 4 | 4 | 19 | 83 | 5 | 5 | 0.0256 |
| 20 | Jordan | 28 | 13 | 12 | 15 | 63 | 6 | 7 | 0.0264 |
| 24 | Palestine | 23 | 7 | 3 | 16 | 98 | 5 | 5 | 0.0268 |
| 8 | Kuwait | 26 | 6 | 5 | 20 | 80 | 6 | 6 | 0.0290 |
| 2 | Kuwait | 19 | 1 | 1 | 18 | 50 | 5 | 6 | 0.0320 |
| 6 | Jordan | 36 | 18 | 5 | 18 | 60 | 5 | 3 | 0.0420 |
| 19 | Syria | 24 | 5 | 5 | 19 | 48 | 5 | 5 | 0.0606 |
| 21 | Kuwait | 24 | 3 | 4 | 21 | 70 | 5 | 4 | 0.1203 |
| 10 | Jordan | 28 | 5 | 6 | 23 | 50 | 4 | 4 | 0.1964 |
| 15 | Sudan | 28 | 3 | 3 | 25 | 55 | 3 | 4 | 0.2116 |
| 1 | Jordan | 36 | 13 | 2.5 | 23 | 62 | 6 | 5 | 0.2244 |
| 16 | Sudan | 34 | 13 | 13 | 21 | 67 | 5 | 3 | 0.2480 |
| 9 | Saudi Ar. | 31 | 12 | 3 | 19 | 60 | 4 | 5 | 0.2737 |
| 23 | Sudan | 30 | 3 | 3 | 27 | 67 | 5 | 4 | 0.2882 |
| 14 | Sudan | 31 | 5 | 5 | 26 | 80 | 3 | 5 | 0.4188 |
| | **MEAN** | **28.1** | **7.7** | **5.7** | **20.4** | **67.1** | **5.0** | **4.9** | **0.1009** |

SJ# = Subject #
YOE = Years subject has spoken English
LOR = Length of residence in an English-speaking region
AOL = Age when subject began to use English to communicate
%USE = Percent daily use of English
SR = Subject's self-rating on pronunciation (1-7)
SR = Experimenter's rating of subject's pronunciation (1-7)
RS = Relative Sensitivity to Temporal Cue (see text)

**Table 2-2: Stimuli in Experiment I**

**Nominal Formant Values (in Hz)**

| Spectral Step | F1 BW=90 | F2 BW=100 | F3 BW=300 |
|---|---|---|---|
| 1 | 233 | 2400 | 3080 |
| 2 | 259 | 2320 | 3016 |
| 3 | 284 | 2240 | 2952 |
| 4 | 310 | 2160 | 2888 |
| 5 | 335 | 2080 | 2824 |
| 6 | 361 | 2000 | 2760 |

**Nominal Characteristics of Formant Transitions**

| Cons. | Formant | Starting/Ending Freq. (Hz) | Duration (ms) |
|---|---|---|---|
| /b/ | F1 | 180 | 25 |
| | F2 | 1465 | 45 |
| | F3 | 2180 | 50 |
| /t/ | F1 | 300 | 40 |
| | F2 | 2000 | 40 |
| | F3 | 2900 | 60 |

**Table 2-3: % Change in 'Beat' ID at each Spectral Step as a Function of Duration**

| Spec. Step | EN Group | AR Group |
|---|---|---|
| 1 | 6 | 51 |
| 2 | 20 | 66 |
| 3 | 31 | 80 |
| 4 | 9 | 59 |
| 5 | 10 | 53 |
| 6 | 3 | 43 |

**Table 2-4: Comparison of Prediction Models (Experiment I)**

| Group | Comp. | ΔG2 | Δdf | Over-disp. | *F* | df | *p* |
|---|---|---|---|---|---|---|---|
| AR | 2 vs 1 | 2485.35 | 1 | 16.746 | 148.42 | 1,33 | 0.000 |
| | 3 vs 1 | 2105.57 | 1 | 16.746 | 125.74 | 1,33 | 0.000 |
| EN | 2 vs 1 | 6519.42 | 1 | 8.060 | 808.87 | 1,33 | 0.000 |
| | 3 vs 1 | 192.24 | 1 | 8.060 | 23.85 | 1,33 | 0.000 |

**Table 2-5: Formant Values of Stimuli in Experiment II**

| Stim. | Cat. | Type | F1 in Hz (mels) | F2 in Hz (mels) |
|---|---|---|---|---|
| 1a | /y/ | Centre | 297 (375) | 1828 (1500) |
| 1b | /y/ | Training | 314 (394) | 1719 (1443) |
| 1c | /y/ | Training | 349 (432) | 1866 (1519) |
| 1d | /y/ | Training | 280 (356) | 1942 (1557) |
| 1e | /y/ | Training | 247 (318) | 1791 (1481) |
| 1f,j | /y/ | Test | 346 (429) | 1780 (1475) |
| 1g,k | /y/ | Test | 322 (402) | 1934 (1553) |
| 1h,l | /y/ | Test | 249 (321) | 1878 (1525) |
| 1i,m | /y/ | Test | 272 (348) | 1726 (1447) |
| 2a | /ø/ | Centre | 366 (450) | 1420 (1275) |
| 2b | /ø/ | Training | 384 (469) | 1326 (1218) |
| 2c | /ø/ | Training | 421 (507) | 1452 (1294) |
| 2d | /ø/ | Training | 348 (431) | 1518 (1332) |
| 2e | /ø/ | Training | 313 (393) | 1388 (1256) |
| 2f,j | /ø/ | Test | 418 (504) | 1378 (1250) |
| 2g,k | /ø/ | Test | 392 (477) | 1511 (1328) |
| 2h,l | /ø/ | Test | 316 (396) | 1462 (1300) |
| 2i,m | /ø/ | Test | 340 (423) | 1333 (1222) |

**Table 2-6: Summary of 'A' and 'B' (%) Scores in Experiment II**

| Group | | A Score | B Score |
|---|---|---|---|
| EN | $\bar{x}$ | 9.2 | 44.1 |
| | RANGE | 0 - 22.5 | 10 - 75 |
| | ST. DEV. | 9 | 17 |
| AR | $\bar{x}$ | 6.1 | 86.9 |
| | RANGE | 0 - 17.5 | 35 - 100 |
| | ST. DEV. | 6 | 18 |

**Table 2-7: Results of Experiment II by Subject (% Duration-matches on 'B' Stimuli)**

**EN Group**

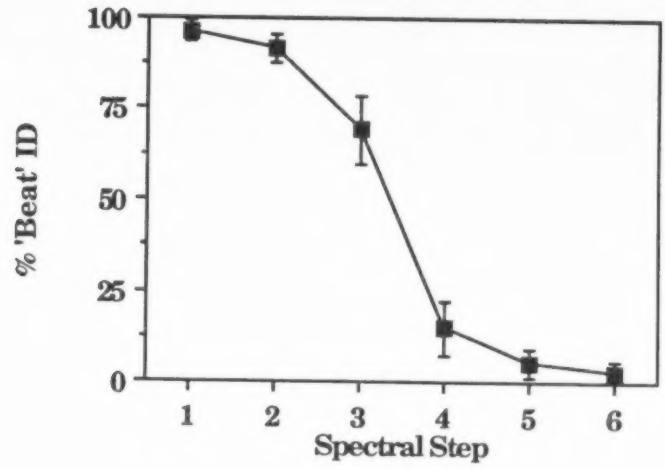
| Subj # | All | /y/ | /ø/ |
|---|---|---|---|
| 1 | 40 | 50 | 30 |
| 2 | 45 | 60 | 30 |
| 3 | 55 | 65 | 45 |
| 5 | 67.5 | 65 | 70 |
| 6 | 55 | 50 | 60 |
| 8 | 32.5 | 60 | 5 |
| 9 | 35 | 35 | 35 |
| 10 | 62.5 | 35 | 90 |
| 11 | 42.5 | 60 | 25 |
| 12 | 10 | 20 | 0 |
| 13 | 20 | 40 | 0 |
| 14 | 42.5 | 20 | 65 |
| 15 | 45 | 35 | 55 |
| 16 | 30 | 30 | 30 |
| 17 | 17.5 | 30 | 5 |
| 18 | 32.5 | 35 | 30 |
| 19 | 67.5 | 85 | 50 |
| 20 | 37.5 | 65 | 10 |
| 22 | 75 | 80 | 70 |
| 23 | 47.5 | 50 | 45 |
| 26 | 42.5 | 35 | 50 |
| 27 | 67.5 | 75 | 60 |
| $\bar{x}$ | 44.1 | 49.1 | 39.1 |
| SD | 17 | 19 | 25 |

**AR Group**

| Subj # | All | /y/ | /ø/ |
|---|---|---|---|
| 1 | 100 | 100 | 100 |
| 2 | 92.5 | 95 | 90 |
| 3 | 77.5 | 70 | 85 |
| 4 | 60 | 50 | 70 |
| 5 | 70 | 50 | 90 |
| 6 | 92.5 | 85 | 100 |
| 7 | 35 | 55 | 15 |
| 8 | 100 | 100 | 100 |
| 9 | 100 | 100 | 100 |
| 10 | 97.5 | 100 | 95 |
| 11 | 100 | 100 | 100 |
| 13 | 52.5 | 65 | 40 |
| 14 | 97.5 | 95 | 100 |
| 15 | 97.5 | 100 | 95 |
| 16 | 97.5 | 95 | 100 |
| 17 | 72.5 | 55 | 90 |
| 18 | 97.5 | 100 | 95 |
| 19 | 95 | 95 | 95 |
| 20 | 92.5 | 85 | 100 |
| 21 | 95 | 95 | 95 |
| 22 | 95 | 100 | 90 |
| 23 | 95 | 95 | 95 |
| $\bar{x}$ | 86.9 | 85.7 | 88.2 |
| SD | 18 | 19 | 21 |

**Figure 2-1: Identification Scores by Spectral Step**

a) EN Group




b) AR Group

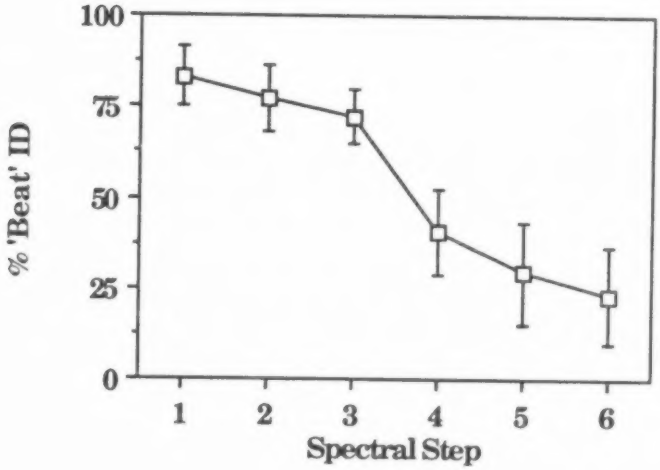
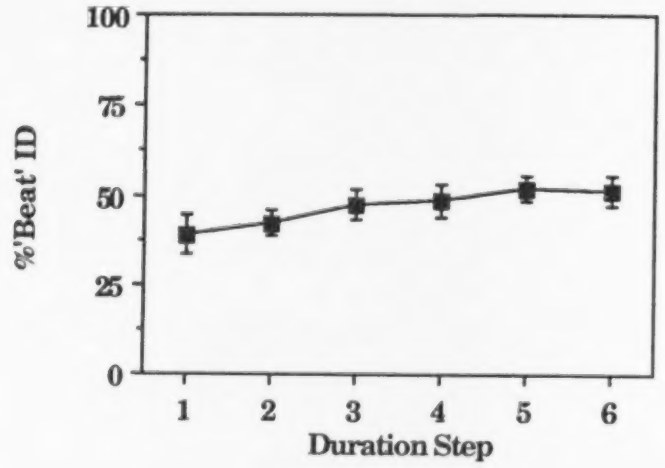


**Figure 2-2: Identification Scores by Duration Step**

a) EN Group




b) AR Group

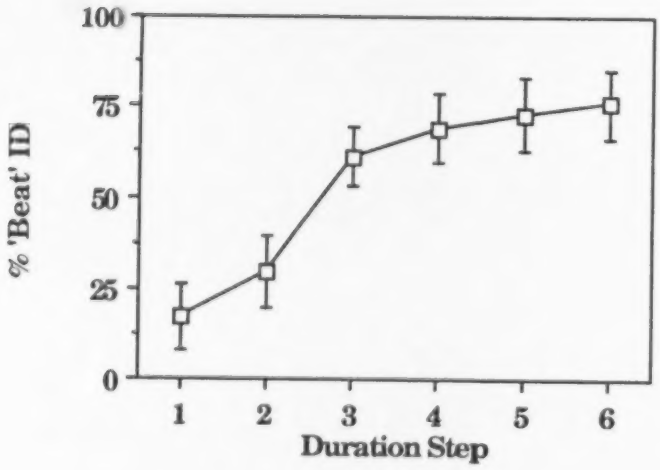
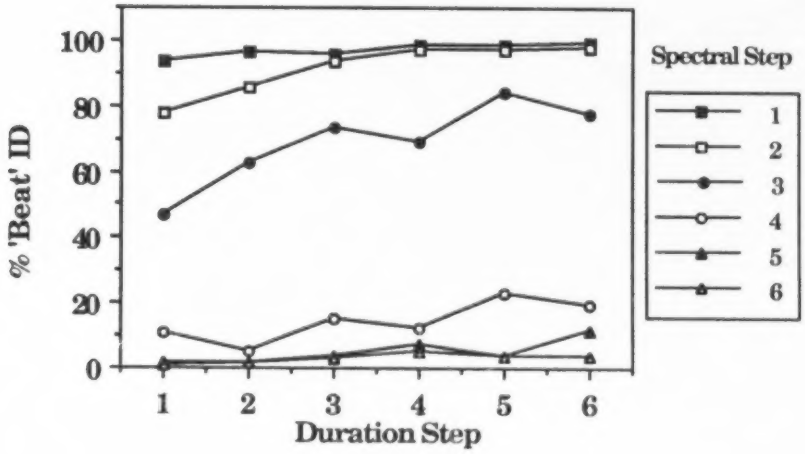


**Figure 2-3: Effect of Spectral Properties at Each Duration Step**

**a) EN Group**




**b) AR Group**

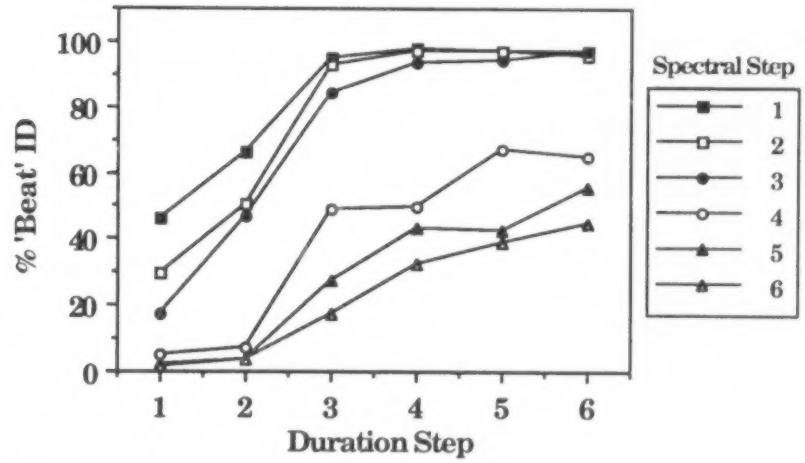


Figure 2-4: Territorial Maps

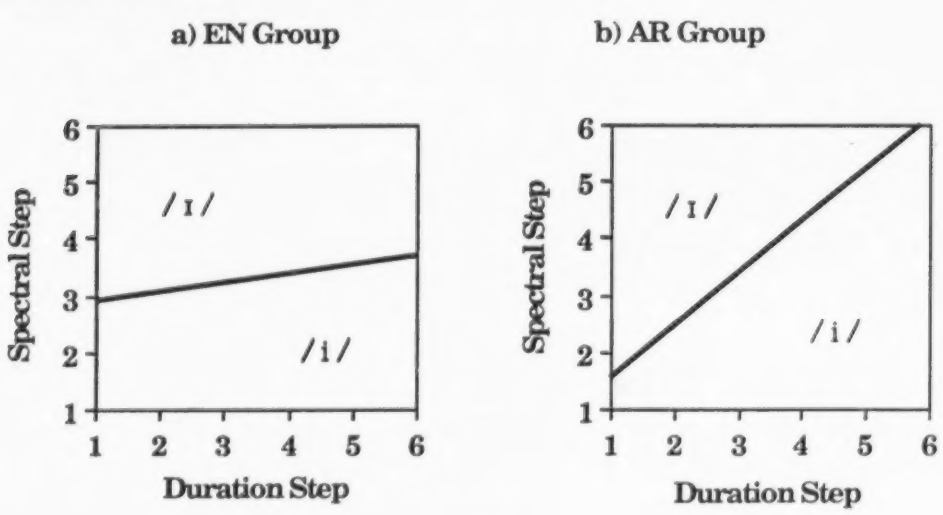



Figure 2-5: Change in % 'Beat' ID Due to Spectral and Temporal Properties

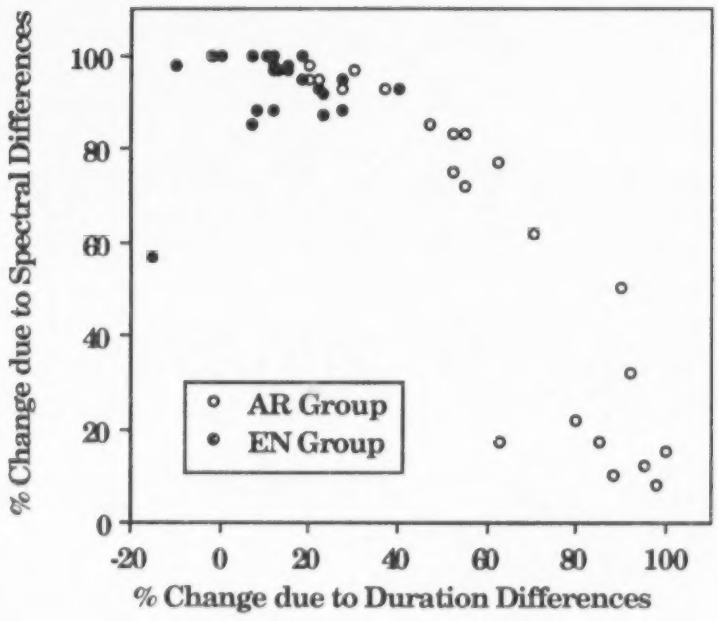


**Figure 2-6: Spectral vs. Temporal Coefficients from the Linear Logistic Analysis**

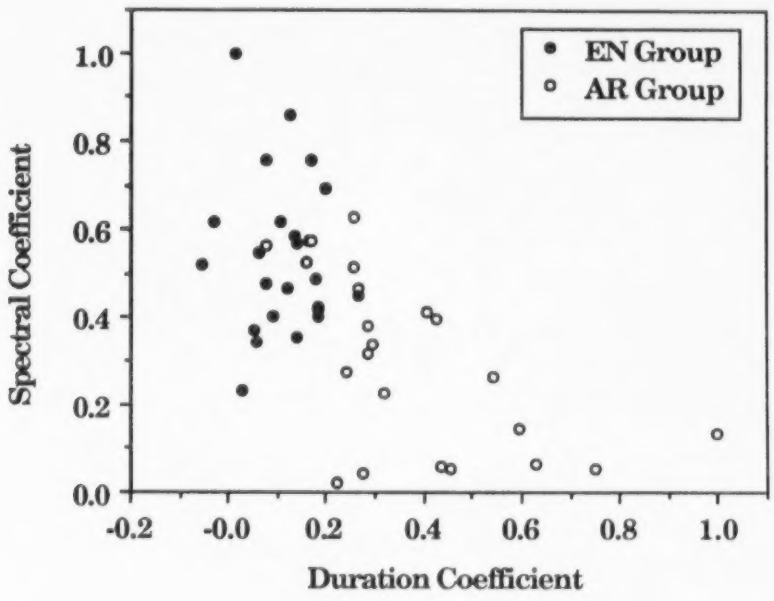
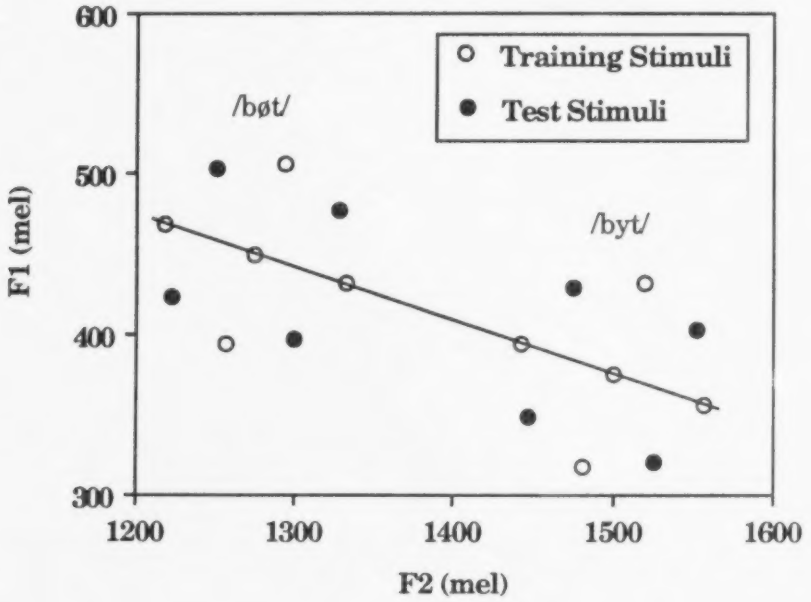


**Figure 2-7: Stimuli in Experiment II**



**References**

Abramson, A., & Ren, N. (1990). Distinctive vowel length: duration vs. spectrum in Thai. *Journal of Phonetics* **18**:79-92.

Ainsworth, W. (1972). Duration as a cue in the recognition of synthetic vowels. *Journal of the Acoustical Society of America* **51**:648-651.

Al-Ani, S. (1970). *Arabic phonology*. The Hague: Mouton.

Bennett, D. (1968). Spectral form and duration as cues in the recognition of English and German vowels. *Language and Speech* **11**: 65-85.

Best, C. (1990). *Adult perception of nonnative contrasts differing in assimilation to native phonological categories*. Paper presented at the meeting of the Acoustical Society of America, San Diego, CA.

Bohn, O-S., & Flege, J. (1990a). Interlingual identification and the role of foreign language experience in L2 vowel perception. *Applied Psycholinguistics* **11**: 303-328.

Bohn, O-S., & Flege, J. (1990b). The role of duration differences in the perception of non-native vowel contrasts. Paper presented at the Annual Meeting of the *Societas Linguistica Europaea*, Berne.

Caramazza, A., Yeni–Komshian, G., Zurif, E., & Carbone, E. (1973). The acquisition of a new phonological contrast: the case of stop consonants in French-English bilinguals. *Journal of the Acoustical Society of America* **54**: 421-428.

Chen, M. (1970). Vowel length variation as a function of the voicing of the consonant environment. *Phonetica* **22**: 129-159.

Crystal, T., & House, A. (1988). The duration of American- English vowels: an overview. *Journal of Phonetics* **16**: 263- 284.

Debrock, M., & Forrez, G. (1976). Analyse mathématique des voyelles orales du néerlandais et du français: méthode et résultats. *Revue de Phonétique Appliquée* **37**:27-73.

Denes, P. (1955). Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America* **27**:761-764.

Elman, J., Diehl, R., & Buchwald, S. (1977). Perceptual switching in bilinguals. *Journal of the Acoustical Society of America* **62**: 971-974.

Ferguson, G., & Takane, Y. (1989). *Statistical Analysis in Psychology and Eduction* (6th ed.) New York: McGraw-Hill.

Flege, J. (1987). The production of "new" and "similar" phones in a foreign language: evidence for the effect of equivalence classification. *Journal of Phonetics* **15**: 47-65.

Flege, J., & Port, R. (1981). Cross-language phonetic interference: Arabic to English. *Language and Speech* **24**: 125-146.

Gottfried, T., & Beddor, P. (1988). Perception of temporal and spectral information in French vowels. *Language and Speech* **31**: 57-75.

Greiser, D., & Kuhl, P. (1989). Categorization of speech by infants: support for speech-sound prototypes. *Developmental Psychology* **25**:577-588.

Hogan, J., & Rozsypal, A. (1980). Evaluation of vowel duration as a cue for the voicing distinction in the following word–final consonant. *Journal of the Acoustical Society of America* **67**: 1764-1771.

Johansson, K. (1984). Perceptual experiments with duration versus spectrum in Swedish vowels. *Lund University Working Papers*, #27.

Jonasson, J., & McAllister, R. (1972). Foreign accent and timing: an instrumental phonetic study. *Papers from the Institute of Linguistics* (PILUS), Publication 14, pp. 11-40. Stockholm: University of Stockholm.

Klatt, D. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America* **59**:1208-1221.

Mitleb, F. (1981). Timing of English vowels spoken with an Arabic accent. *Research in Phonetics*, Report No. 2, pp. 193-226. Bloomington:Indiana University, Department of Linguistics.

Nearey, T. (1990). The segment as a unit of speech perception. *Journal of Phonetics* **18**: 347-373.

Norlin, K. (1984). Acoustic analysis of vowels and diphtongs (sic) in Cairo Arabic. *Lund University Working Papers*, #27.

Peterson, G. & Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America* **32**: 693-703.

Raphael, L. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *Journal of the Acoustical Society of America* **51**: 1296-1303.

Umeda, N. (1975). Vowel duration in American English. *Journal of the Acoustical Society of America* **58**:434-445.

Underhill, R. (1976). *Turkish grammar*. Cambridge: The MIT Press.

Van Heuven, V. (1986). Some acoustic characteristics and perceptual consequences of foreign accent in Dutch spoken by Turkish immigrant workers. *Papers Presented at the Dutch Linguistics Colloquium*, Berkeley. J. van Oosten and J. Snapper (Eds). Berkeley: The Dutch Studies Program.

Wiik, K. (1965). Finnish and English Vowels. Turku: Turun Yliopiston Julkaisuja Annales Universitatis Turkuensis.

Winer, B. (1971). *Statistical principles in experimental design* (2d ed.). New York: McGraw-Hill.

# CHAPTER 3
# EXPERIMENT III: ENGLISH VOWEL PRODUCTION BY NATIVE SPEAKERS OF ARABIC

## Introduction

Experiments I and II revealed differences in the ways native speakers of English and native speakers of Arabic attend to spectral and temporal properties of certain vowel pairs. Experiment III was carried out in an effort to explore the differences in the English vowel *productions* of the same two groups of speakers.

Flege (1987a) discusses some of the methodological issues which arise in the collection and interpretation of L2 production data. He observes that instrumental studies of pronunciation by non-native speakers generally endorse the notion that among native speakers, phonetic norms exist for various acoustic properties of speech, such as voice onset time in stops and formant frequencies in vowels. The productions of L2 learners may be compared against these norms in an effort to establish how native-like or "authentic" such productions are. A number of researchers have compared native and non-native productions in terms of a variety of acoustic properties. For instance, numerous studies have examined differences in voice onset time (Caramazza, Yeni-Komshian, Zurif, & Carbone; 1973; Flege, 1980, 1987b, Williams, 1980). These studies have shown that speakers often do not produce native-like VOT intervals in their L2, even if they have extensive experience with it. Studies of vowel production (Mitleb, 1981; Port & Mitleb, 1983; Jonasson & McAllister, 1972) have also shown effects of L1 on the production of vowel duration differences in L2. Some recent studies have shown that L2 learners have varying degrees of success in producing L2 vowels with native-like spectral properties (Flege & Hillenbrand, 1984; Flege, 1987b).

Much of the previous work on non-native speech production has tested predictions regarding which vowels and consonants ought

to be produced most accurately by L2 learners, given the inventories of L1 and L2. Several studies by Flege and his colleagues (Flege, 1987b; Flege, 1988; Bohn & Flege, 1990), for instance, have tested the proposal that "new" sounds are more readily acquired by L2 learners than sounds which are "similar" to sounds in the L1. In the present research, the ways in which the L1 sound system influences vowel production in L2 will also be considered. However, rather than focus intensely on the production of one or two vowels, this study will examine more broadly how well a group of Arabic speakers who have learned English in adulthood succeed in producing a relatively wide range of English vowels. Their productions will be compared with those of native speakers with respect to a number of temporal and spectral characteristics.

***Vowel Duration***

Mitleb (1981) examined vowel production in seven native speakers of Jordanian Arabic and seven native speakers of English. The Jordanians produced the six vowel phonemes /i:/, /i/, /ɑ:/, /ɑ/, /u:/, and /u/ from their native dialect, as well as a number of English vowels, in /bVt/ and /bVd/ contexts within a sentence frame. He reported a long–short duration ratio for Jordanian vowels of 1.5. Apparently because they partially transferred Arabic vowel duration patterns to English, the Jordanians exaggerated vowel duration differences in English tense-lax pairs, producing a tense-lax duration ratio of 1.3 - smaller than the long-short ratio in Arabic, but significantly larger than the tense-lax ratio of 1.2 produced by the native English speakers. He also observed that overall vowel durations were shortest for the Jordanians' Arabic productions, longest for the English speakers' English productions, and intermediate for the Jordanians' English productions.

In the Jordanian vowels examined in Mitleb's study, a small effect of consonant voicing on preceding vowel duration was found.

Flege & Port (1981) reported a similar finding based on data from six Saudi speakers. When producing English monosyllables, both the native English and the Jordanian speakers in the Mitleb study showed a significant effect of consonant voicing on vowel duration. However, while the English group produced a long–to–short vowel ratio of 1.3 the Jordanians produced a smaller ratio of less than 1.1.

Experiment III will replicate Mitleb's assessment of Arabic speakers' productions of English vowel durations. Here, a larger group of subjects representing several dialects will be used. Performance on English tense-lax vowel pairs and on vowels preceding /t/ and /d/ will be assessed. Also, some attention will be given to individual differences in production. One hypothesis to be tested concerns the relationship between the results obtained here and the perceptual data from Experiment I. In that study, it was found that the subjects exhibited large differences in the use of temporal and spectral cues in the identification of vowels from an /i/–/ɪ/ continuum. It might be expected that those subjects who showed the most English-like patterns of identification should produce these vowels with a duration ratio relatively close to that of native speakers.

***Vowel Quality***

While the vowel systems of all Arabic dialects show contrastive length, there is some dialectal variation in inventories. Table 3-1 summarizes these differences for most of the dialects represented in the subject sample in this study. "Standard" Arabic is analysed as having three basic qualities (/i/, /ɑ/, /u/), while regional dialects usually make more distinctions. Holes (1990) describes Gulf Arabic (including the Arabic spoken in Kuwait and Eastern Saudi Arabia) as having five long (/iː/, /eː/, /ɑː/, /uː/, /oː/) and three short (/i/, /ɑ/, /u/) vowels. According to his transcriptions, his observation that the long vowels are "tense" and the short vowels are "lax" means

that /i/ is usually realized as [ɪ], /u/ as [ʊ], and /ɑ/ as [æ], while the long vowels are [i: ], [e:], [æ: ], [u:], and [o:]. Gary & Gamal-Eldin (1984) indicate that Cairene Egyptian Arabic has a very similar vowel system to that of Gulf Arabic, again with the same five long vowels and three short ones. They also state that the long vowels are tense and the short ones lax. According to Cowell (1964), Syrian Arabic has the five vowel qualities discussed above, as well as /ə/. The first five all have phonemic long and short forms. His description indicates that the long vowels and their short counterparts are similar in quality, but that the short ones are "less tense." Trimingham (1959) reports the same inventory for Sudanese Arabic. One additional feature of Arabic vowels which is worthy of note is that, regardless of dialect, they typically show much less diphthongization than English vowels (Mitchell, 1990).

Although some studies have compared formant frequencies of vowels produced by native and non-native speakers (Flege & Hillenbrand, 1984; Flege, 1987b), the research comparing productions of several vowels on multiple spectral dimensions is fairly limited. Most studies have relied on single measurements of F1 and F2, usually taken at vowel midpoints, as a means of quantifying accentedness in L2 productions. One serious drawback to the use of single data points as estimates of formant frequencies is that it does not permit an assessment of formant movement. Since diphthongization is an important aspect of English vowel production, an evaluation of English vowels produced by non-natives ought to include an analysis of formant movement data. In Experiment III, F1 and F2 values of ten English vowels produced by native Arabic and native English speakers will be measured at two points, one relatively early in the vowel and one relatively late. The differences between these two sets of measurements will be used to estimate formant movement. The measurement data, then, will be used to compare

the productions of the two groups on four dimensions: F1 and F2 frequencies, and F1 and F2 movement.

The difficulties which arise in making such a comparison of spectral properties are similar to those which face researchers wishing to compare vowels or vowel systems from different languages or dialects. For instance, variation due to individual production differences, arising primarily from differences in vocal tract length, must be controlled for. Hindle (1978) compared three normalization procedures in a correlational analysis of dialectal change in a single vowel produced by both male and female speakers of varying ages. The log mean transformation due to Nearey (1977) proved to be especially successful in revealing age-related articulation effects without over-normalizing the data. Considerable variability is removed from formant frequency data if subjects of only one sex and age group (e.g all adults) are considered. Disner (1986) used analysis of variance in a comparison of untransformed English and Dutch F1 and F2 frequencies of vowels produced by adult male speakers. Holden and Nearey (1986), however, applied a log transformation prior to using ANOVA in their analysis of differences between the vowels of male and female speakers of three dialects of Russian.

The information available on the Arabic vowel quality leads to some questions about how speakers of Arabic will produce English vowels. For instance, the Arabic vowels /i:/, /i/, /u:/, /u/, and (in some dialects) /e:/, /e/, and /o:/ appear to correspond to the English vowel categories /i/, /ɪ/, /u/, /ʊ/, /e/, /ɛ/, and /o/. In the terms used by Flege (see above) the English vowels are most likely "similar" to the L1 vowels and are therefore unlikely to be produced in a native-like way because of the effects of equivalence classification. A good exemplar of any of these vowels from one of the Arabic categories would not necessarily be a good exemplar of the corresponding English category. In the first place, there are likely to

be differences in formant values between the two languages. Furthermore, as noted previously, Arabic does not show the same degree of formant movement that English does. Nearey & Assmann (1986) observed significant movement not only in the traditional English diphthongs /e/ and /o/, but also in /ɪ/, /ɛ/, and /æ/ when produced in isolation. Andruski & Nearey (in press) reported a similar finding. In order to produce these and the other vowels correctly in English, a speaker of Arabic would have to note and produce the sometimes subtle differences between the L1 and L2 categories. One question to be answered here, then, is whether the speakers of Arabic will succeed in producing these vowels with spectral properties similar to those of native speakers.

Even though all dialects of Arabic have the vowel /a:/ , it is rather difficult to predict how well the subjects in this Experiment will produce English /ɑ/. Arabic /a:/ is often realized as a relatively front vowel, but apparently varies in quality from dialect to dialect. One might therefore expect to see differences in how speakers of different dialects produce English /ɑ/.

Several general questions will be considered in this study. For instance, do the Arabic speakers as a group show native-like production of any of the English vowels? What individual differences exist in how some of the vowels are produced? Do vowels produced by Arabic speakers show appropriate patterns of formant movement?

## Methods

### *Subjects*

Subjects were 23 native speakers of Arabic and 23 native speakers of English. The 21 male and 2 female Arabic speakers had all participated in Experiment I. For a complete list according to dialect, see Table 2–1 in Chapter 2. One of the female subjects from the native English group in Experiment I was dropped, and an

additional male was recruited, for a total of 12 male and 11 female speakers.

***Materials and Procedure***

Recordings were made on a Technics M235X cassette recorder with a Nakamichi CM300 microphone. Subjects were seated in a sound–treated room with their lips approximately 15 cm from the microphone. Materials were similar to those used in the Mitleb (1981) study. The subjects were asked to read a list of /bVt/ and /bVd/ words in the sentence frame "I said _____ and then left the room." The words were chosen to elicit the ten vowels /i/, /ɪ/, /e/, /ɛ/, /æ/, /u/, /ʊ/, /o/, /ɑ/ and /ʌ/, and were written in standard orthography. In the two cases in which the desired token was not a real word, a "sounds–like" example was provided (i.e., in the case of /bʊt/ the word *put* was listed, and to elicit /bʊd/, the word *good* was listed). To elicit /bɑt/ the spelling used was *bot* . To prevent a list effect, two additional sentences were added at the end of the list. The subjects were encouraged to give their best possible production of each word. If the experimenter suspected that an error in production was simply a reading error, rather than a genuine indication of the subject's inability to produce a particular vowel correctly, the subject was asked to repeat the entire sentence.

The 920 tokens were digitized at 16.7 kHz with CSRE software (Jamieson, Ramji, Nearey, & Baxter, 1990) in the phonetics laboratory at the University of Alberta. The edited tokens were then submitted to LPC analysis for later formant measurements. A 15 ms window was used with a 2 ms hop and twelve coefficients.

Measurements were made with a mouse from a Macintosh computer screen display giving a time–domain representation as well as formant tracks from the LPC analysis for F1 to about F4. Vowel duration was measured to the nearest 0.1 ms with cursors positioned on the waveform at the first sign of periodicity after the

release of the initial /b/ and at the end of the vowel. The latter point was marked either by the beginning of closure voicing, which was evidenced as a drop in the amplitude of the signal, a change in the shape of the waveform, and a sudden drop in F1 frequency, along with the disappearance of higher formants; or, when no closure voicing was present, by the beginning of a silent interval.

Formant values for F1 and F2 were obtained by visually positioning the cursor on the formant tracks. Since a single measurement of a formant does not characterize formant movement, a factor which might be relevant in how well the AR group produced the English vowels, it was decided that two measurements would be taken from each formant. Measurements were therefore made at approximately 30% of the distance after the beginning of the vowel (the *a* measurement) and 30% of the distance before the end (the *b* measurement).

## Results

### *Duration Data*

The measured vowel durations were transformed with a log transformation to help control for inter–subject variability (due, for instance, to differences in speaking rate) and were submitted to a three–way repeated measures analysis of variance (Native Language X Vowel X Final Consonant). The analysis revealed significant main effects of all three factors, as well as significant LV ($F$ (9,396)=33.08; $p < 0.0001$) and LC ($F$ (1,44)=142.34; $p < 0.0001$) interactions. The VC and LVC interactions proved to be non–significant ($F$ (9,396)=1.11 and $F$ (9,396)=1.80 respectively).

### *LV Interaction*

The mean durations by vowel (untransformed) for the two groups are given in Table 3-2 and illustrated in Figure 3-1 and 3-2. In the figures, the vowels are grouped according to the categories *front* and

*back* . The front vowel data showed a similar general pattern of results for both groups: /æ > e > i > ɛ > ɪ/ . Post-hoc tests using the Tukey (a) procedure (Winer, 1971) revealed that in the EN group all differences were significant at at least the 0.05 level. This finding agrees with previous studies which generally show low vowels to be longer than high ones and tense vowels to be longer than lax ones (Umeda 1975; Crystal & House, 1988). In the AR group the /æ/– /e/ difference failed to reach significance, but all other differences were significant at the 0.01 level.

In their productions of the back vowels, the EN group showed the general pattern /ɑ > o > u > ʌ > ʊ /. Again, all differences were significant at at least the 0.05 level. The AR group differed somewhat, showing the ordering /o > u > ɑ > ʊ > ʌ/ . While the /ɑ/–/ʊ/ difference was not significant, all other differences were significant at at least the 0.05 level. The most notable difference between the two groups here was in the duration of /ɑ/. While this vowel was significantly longer than all other back vowels in the EN data, the AR subjects produced it as considerably shorter than /o/ and /u/.

In order to determine whether the AR group exaggerated duration differences between tense and lax vowels, *t* –tests were performed on the tense–lax vowel duration ratios from the two groups on the three vowel pairs /i/–/ɪ/, /e/–/ɛ/, and /u/–/ʊ/. The tense–lax duration ratios in the AR group were 1.7, 1.7, and 1.5 respectively. In the EN group the ratio was 1.2 in all three cases. The between–group differences were highly significant on /i/–/ɪ/ ($t$ (44) = 5.592; $p$ < 0.0001), /e/–/ɛ/ ($t$ (44)=8.276; $p$ < 0.0001), and /u/–/ʊ/ ($t$ (44)=3.76; $p$ < 0.0005), indicating that the AR group did indeed produce a significantly larger duration contrast in these vowel pairs than did the EN group.

When vowel durations were compared across the two groups, it was found that the EN speakers produced every vowel except /u/

with significantly greater duration than did the AR speakers ($p < 0.01$ in all cases). An important question which arises here is whether the shorter vowel durations in the AR group might be due to a faster speaking rate among the AR subjects. It might be hypothesized, for instance, that L2 learners are more likely than native speakers to experience anxiety during a production task, and might therefore speak more quickly. To test this proposal, the carrier sentences containing the words /bæt/ and /bæd/ were digitized and their durations measured to the nearest 1 ms. These sentences were chosen first because they were in approximately the middle of the /bVt/ and /bVd/ lists and should therefore be fairly representative of the speaking rate used by speakers throughout each list and second because all subjects' productions of these sentences were free of hesitations which might lead to exaggerated durations in the measurement data. Two-sample *t*-tests revealed no significant differences in duration in the case of either the /bæt/ sentences[1] ($t(22)=1.598$; two-tailed) or the /bæd/ sentences ($t(22)=0.183$). There is no indication, then, that the vowel duration differences observed here were due to different speaking rates in the two groups.

*LC Interaction*

Table 3-3 and Figure 3-3 give the mean durations pooled over vowels before each consonant. The mean durations before /d/ and /t/ in the EN group were 277.3 and 192.2 ms respectively, giving a ratio of over 1.4 to 1. In the AR group the durations were 188.7 and 179.6, giving a ratio of less than 1.1 to 1. Post hoc comparisons with the Tukey (a) procedure revealed that both the EN ($p < 0.01$) and the AR groups ($p < 0.01$) produced significantly longer vowels before /d/ than before /t/. However, further exploration of these data revealed that the EN

[1]In fact, the AR speakers' productions of the /bæt/ sentences were actually an average of 216 ms longer than EN speakers' productions.

group produced a significantly larger long-to-short vowel duration ratio than did the AR group ($t$ (44) = 10.898, $p$ < 0.0001).

The vowels produced by the EN group proved to be longer in both consonantal contexts than the vowels produced by the AR group ($p$ < 0.01). In addition, the duration of vowels before /t/ in the English group was actually longer than that of vowels before /d/ in the AR group ($p$ < 0.01).

***Spectral Data***

An analysis of the spectral properties of the two groups' productions is a more difficult undertaking than is an examination of durational properties. More than one dimension must be examined, since vowels can differ in F1, F2, and higher formants, as well as in the extent and direction of movement of those formants. In addition, formant frequencies are subject to large effects of speaker. To reduce some of this variability, data from only the male speakers were analyzed (21 of the 23 original AR speakers and 12 of the original 23 EN speakers). The formant values were averaged over the two replications from each speaker prior to further analysis. Table 3-4 gives the means in Hz of the F1a and F2a measurements from both groups. These values have also been plotted in Figure 3-4. To indicate how the productions of the two groups differ, the mean formant values of the AR group (symbolized by dark squares) and those of the EN group (hollow circles) have been connected.

Also of interest here is the amount and direction of formant movement in the productions of the two groups. Table 3-5 gives the means in Hz of the $b$ - $a$ values of F1 and F2 (labelled ΔF1 and ΔF2) for both groups. Positive values indicate rising formant frequencies, while negative values indicate falling frequencies. These differences are also represented in Figures 3-5 and 3-6. In their study of Western Canadian English vowels produced in isolation, Nearey & Assmann (1986) (NA) observed significant downward movement in F1 of /e/

and /o/ and upward movement in /ɪ/ and /ɛ/. In F2 they reported significant downward movement in /ɪ/, /ɛ/, /æ/, and /o/, and upward movement in /e/. As can be seen from Table 3-5 all these patterns were observed in the English speakers in the present study, according to single sample *t* -tests (two-tailed). In addition, in several other cases formant movement was found to be significant. In F1, significant upward movement was noted in /æ/, /ʊ/, and /ɑ/, and downward movement was seen in /u/. In F2, significant upward movement occurred in /i/, /ʊ/, /ɑ/, and /ʌ/. With the exception of the rise in F2 of /i/, all these observations agree with trends observed by NA. The slight differences between these two studies might be accounted for first by the fact that different dialects were considered and second because NA examined vowels in isolation, while the vowels studied here were produced in /bVt/ and /bVd/ contexts.

An examination of the AR speakers' data revealed that they failed to show significant movement in several cases where the EN speakers did. In only one case (F2 of /u/), did the AR group show significant movement where the EN group did not.

To explore the differences between the two groups further, the measurements were transformed with a log transformation (Holden & Nearey, 1986), and *t* - tests were performed on four sets of data: log(F1a), log(F2a), log(F1a)-log(F1b), and log(F2a)-log(F2b). Table 3-6 gives the *t* - values and significance levels for the tests on each vowel. An examination of the front vowel data reveals a significant difference in the F1a value only for the vowel /ɛ/, which was due to a lower F1 value in the AR group. No differences were observed on the F2a values of the front vowels. Significant differences emerged on every vowel in the F1 movement data and on every vowel except /i/ in the F2 movement data. In general then, the AR speakers seemed to achieve native–like values of F1 and F2 at an early point in the production of all front vowels except /ɛ/, but they failed to show

native-like formant movement for the full duration of the vowel. In general this was because they did not exhibit the same *degree* of movement as did the EN group, but, as can be seen from Figures 3-5 and 3-6, in the case of F1 of /i/ and F2 of /ɪ/, the AR group actually showed a small amount of movement in the direction opposite that of the EN group. In all other front vowels, formant movement was in the same direction for both groups.

An examination of the back vowels revealed that the two groups differed on the F1a measurements of the vowels /ɑ/ and /ʌ/. In the case of /ɑ/ this difference was due to a considerably lower F1 value (almost 100 Hz lower) in the AR group, while for /ʌ/ the AR speakers produced a higher value of F1. As can be seen from Figure 3-4, the effect of this difference, combined with a lower value of F2 in the AR group, is to bring these two vowels much closer together in the F1-F2 space than the corresponding native productions. The two groups differed on the F2a measurements of all the back vowels except /ɑ/. In all cases, this was because the AR speakers produced lower F2 values. In general, then, the AR group obtained less accurate F2 values, with respect to the EN group, on the back vowels than on the front ones. They also exhibited differences in F1 movement in /u/, /ʊ/, and /o/ and F2 movement in /u/ and /o/. Once again, these differences are mostly due to degree of movement, but in the case of /u/, the movement is slightly in the opposite direction for F1 and somewhat more strongly in the opposite direction for F2.

In order to rule out the possibility that the differences in formant values described here are due in some way to systematic differences in vocal tract length, F3 values were obtained from the subjects on the vowels /ɛ/, /æ/, and /ʌ/ in the /bVt/ context. Measurements were made in the same manner as for F1 and F2, except that only one measurement was made, this time at approximately the midpoint of each vowel. The mean of the three

vowel measurements from each subject was taken and submitted to a two-tailed *t* - test which indicated no significant difference ($t$ (31) = 0.709, $p$ = 0.4836). Since there was no substantial difference in F3 frequencies, it appears that the between-group differences are due to differences in the ways the two groups articulated the vowels.

***Individual Differences in Production***

No attempt will be made here to consider all the possible ways that individual Arabic subjects might have differed from each other in the production of the English vowels. A few key differences will be explored, however.

Table 3-7 gives the durations of the individual subjects' productions of three tense-lax vowel pairs in the /bVt/ context. Also shown are the tense-lax duration ratios for these pairs. It is clear that some subjects came much closer to producing native-like ratios than others. On the /i/–/ɪ/ pair, eight subjects showed fairly native-like performance, using the criterion of ±1 standard deviation (0.28) from the English mean of 1.2 as the range within which performance might be considered native-like. On the /e/–/ɛ/ pair only two subjects fell within one standard deviation (0.12) of the English mean of 1.1, and on the /u/–/ʊ/ pair, six subjects showed native-like ratios using the English mean of 1.1 (s.d.=0.18) as the basis for comparison. Subjects who fell within the native range on one vowel pair did not necessarily do so on the other pairs, and only one subject did so on all three pairs. This subject ranked eleventh in terms of size of the relative sensitivity score from Experiment I (see Table 2-1), which indicates that he did not show an especially native-like pattern of results in the perceptual study.

Table 3-8 gives the differences in Hz between the F1a and F2a values of the vowels /i/ and /ɪ/ produced by the AR subjects and the frequencies of these formants in the EN data. These figures are absolute values and therefore do not indicate direction (see next

section). Once again using the criterion of ±1 standard deviation from the English mean as the standard for native-like performance, only two subjects succeeded in producing native-like formant values in all four cases. In general, the subjects showed a wide range of variation in the accuracy of their F1 and F2 frequencies, with some producing F2 values more than 200 Hz from the English mean.

Table 3-9 illustrates some further individual differences on three vowels of special interest. As shown in Table 3-6, the AR and EN groups differed significantly in amount of F1 movement in /e/, and F1a frequencies in /ɛ/ and /ɑ/. In the measurements of /e/ in the EN group, F1 fell by as little as 47 Hz and as much as 111 Hz. Table 3-9 gives the amount of change (in Hz) in the individual Arabic subjects' productions of this vowel. The subjects have been arranged according to region of origin so that any obvious dialectal differences might be observed. However, because of the small numbers of subjects representing each dialect, it is unlikely that any firm conclusions about dialectal differences can be drawn here. It can be seen that three subjects actually showed F1 movement in the wrong direction, while several others showed only a small amount of movement. An examination of the subjects within dialect regions reveals that in all cases, except for the two Syrian subjects, both large and small amounts of movement were observed.

The second column of Table 3-9 gives the differences (in Hz) between the F1 frequency of /ɛ/ measured from each subject and the English mean value of 521 Hz (SD=31). Every subject but one produced this vowel with F1 lower than the English mean, and several differed from the English mean by much more than the English group's standard deviation of 31. While the Syrian and Sudanese varieties of Arabic have a contrast between /e:/ and /e/, there is no clear evidence here that the subjects who spoke these dialects performed differently as a group from the other subjects who spoke dialects without such a contrast.

The individual differences on F1 values of /ɑ/ are also given in Table 3-9. Every subject produced this vowel with an F1 value lower than the English mean, and all but two values differed from this mean by more than the English group's standard deviation of 33. The relatively large differences in all dialect groups again argue against dialectal effects in the data.

***Relating the Perception and Production Data***

It might be expected that Arabic-speaking subjects who perceive English vowels in a manner similar to native speakers will tend to produce them more accurately (in terms of native productions) than will those whose perceptions are less native-like. In order to test this hypothesis, the data from Experiments I and III were compared in two ways.

First, the relative sensitivity ratios calculated in Experiment I (see Table 2-1) were compared with the duration data presented here. It was expected that subjects showing a large degree of sensitivity to duration when labelling the /i/ - /ɪ/ continuum would tend to exaggerate the duration difference between /i/ and /ɪ/ more than the other subjects. In fact, a Pearson correlation coefficient of 0.579 ($df = 21, p < .01$) was computed between the relative sensitivity values from Experiment I and the corresponding ratios of /i/ to /ɪ/ durations produced by the subjects in the /bVt/ context. (See Table 3-7). While at first this moderate correlation may appear to indicate a fairly consistent relationship between the two scores, a closer inspection of the data suggests that this is not the case. Because of the relatively small sample size, it appears the data from subject 14, who showed a very high degree of relative sensitivity to duration and a very large /i/ to /ɪ/ duration ratio, may have resulted in a distorted estimate of the strength of the relationship. In fact, when this data point is omitted from the analysis, the correlation drops to a non-significant value of 0.26. However, the results do show a trend in

the expected direction. Furthermore, of the eight subjects who exhibited relative sensitivity values greater than 0.1 in Experiment I (i.e., those who relied most on duration in their judgments) only two fell within the native range on the /i/-/ɪ/ pair. This suggests that while native-like perception of duration does not guarantee native-like production, subjects who show extreme duration use in perception are less likely to succeed in producing a native-like duration contrast.

A second way to examine the relationship between the results of the perception and production studies is to compare the spectral properties of the AR groups' productions with the relative sensitivity values from Experiment I. It might be expected that those subjects who produce /i/ and /ɪ/ with spectral values close to the mean values of the English subjects might be the same ones who attend most to the spectral properties of the stimuli (i.e. those who show the smallest relative sensitivity values in Experiment I). To test this prediction, a simple technique, similar to the approach described in Chapter 4, was used. The mean values of F1, ΔF1, F2, and ΔF2 from the /i/ and /ɪ/ productions in the EN data were subtracted from the corresponding values from the AR data, and the absolute values of the resulting differences were taken. The new set of F1 and F2 differences (see Table 3-8) represents the differences between individual tokens from the AR group and the English means along two dimensions, without regard to direction. The ΔF1 and ΔF2 differences represent differences in formant movement. Pearson correlation coefficients were computed between the two sets of differences (one from each vowel) and the relative sensitivity values calculated in Experiment I. The results of this analysis are summarized in Table 3-10. It can be seen that there are indeed significant correlations between some of the measurement data and the perceptual data. The highest correlation was with the F2 difference score for /ɪ/ ($r = 0.638$; $p < 0.01$). The value of the ΔF1

difference for /i/ was also moderately correlated ($r = 0.43$; $p < 0.05$). Both of these correlations were in the expected direction. That is, the greater the difference between the values on an individual Arabic speaker's token and the English mean values on these dimensions, the greater (and therefore less native-like) was the Arabic speaker's use of duration in Experiment I. However, the F1 difference data from the vowel /i/ showed a correlation in the direction opposite the one expected ($r = -0.467$; $p < 0.05$). In other words, the subjects who differed most from the native speakers on this dimension showed the most native-like performance in Experiment I.

## Discussion

### *Vowel Duration*

The duration data in the present experiment conform fairly closely to data from previous studies which indicate that Arabic speakers of English differ from native English speakers in the temporal aspects of English vowel production. The AR speakers produced the same ordering of vowel duration differences for the front vowels as did the EN speakers (/æ > e > i > ɛ > ɪ/), but a different ordering for the back vowels because they produced /ɑ/ as much shorter than did the EN group. As was the case in Mitleb's (1981) study, the magnitudes of the differences between vowels differed in the two groups because the Arabic speakers produced larger vowel duration differences in tense–lax pairs than did the native English speakers. While the EN group produced relatively small tense-lax duration ratios in the /i/-/ɪ/, /u/-/ʊ/ and /e/-/ɛ/ pairs, the AR group produced exaggerated differences relative to the EN group in all three cases. This was apparently because they produced very short lax vowels.

These findings indicate the influence of L1 on the production of L2 vowels. The results of Experiment I (see Chapter 2) have already shown that many Arabic speakers tend to use temporal properties

more than native English speakers when labelling vowels from an English /i/–/ɪ/ continuum. Although only weak evidence could be found here that subjects tend to produce an exaggerated duration difference between /i/ and /ɪ/ contrast *in proportion* to how strongly they hear this contrast in terms of duration, the fact remains that the majority of the Arabic speakers neither heard nor produced the temporal characteristics of this pair in the same way that the native speakers did. It seems reasonable to suppose, then, that the exaggerated use of duration in this production study is rooted in perception. That is, Arabic speakers do not produce English tense-lax pairs in the same way that native speakers do because they do not perceive them in a native-like manner.

The production patterns observed here may in part be the result of substituting Arabic long and short vowels for English tense and lax ones. The evidence from sources cited earlier suggests that Arabic long and short vowel categories may be similar to English tense and lax categories in terms of quality and duration. Therefore, because of equivalence classification (Flege, 1987b), an Arabic speaker may hear English vowels in terms of Arabic categories, and, in production, may substitute Arabic /i:/, /i/, /u:/, and /u/ for English /i/, /ɪ/, /u/, and /ʊ/. Because the Arabic long-short pairs exhibit a greater duration difference than the English tense-lax pairs, however, it might be expected that the Arabic speakers' productions of English vowels will show exaggerated duration effects. In the data presented here these effects are most noticeable in the very short productions of the English lax vowels.

However, some of the results reported here are not immediately explained by this account. For instance, one might expect the Arabic speakers to identify English /ɑ/ with long Arabic /ɑ:/, but this was apparently not the case. Instead, they produced /ɑ/ with an inappropriately short duration, which suggests that they may have associated English /ɑ/ with Arabic short /ɑ/. In fact

Arabic /ɑ:/ and /ɑ/ exhibit a great deal of variation in quality, both cross-dialectally and within dialects. Depending on phonetic context, these vowels may be similar to English /ɛ /, /æ/, /ɑ/, or /ʌ/. It appears that Arabic long /ɑ:/ is most commonly realized as a front vowel similar to /æ/ (Holes, 1990; Mitchell, 1990; Trimingham, 1959) in the dialects represented here. It may be because of a difference in quality, then, that the subjects did not substitute Arabic /ɑ:/ for English /ɑ/. An informal evaluation by the experimenter indicated that many of these productions sounded much like English /ʌ /, which may indicate that the subjects had identified English /ɑ/ with the [ʌ ] variant of short Arabic /ɑ/. Furthermore, Figure 3-4 shows that although the Arabic speakers did produce English /ɑ/ and /ʌ/ at slightly different locations in the F1-F2 space, they did not distinguish the quality of these vowels as much as the native speakers did.

It should also be noted that some of the dialects represented in the sample do not have a phonemic contrast between /e/ and /ɛ/, and some do not have either of these vowels, or the vowel /o/. However, /o/ was produced as only slightly shorter than native /o/, and /e/ and /ɛ/ were produced with an exaggerated duration difference by all but two of the subjects (see Table 3-7). There was no evidence that dialect played a role in the production of duration differences between these two vowels. In the case of those subjects whose dialect does not distinguish this pair it is not clear why the subjects produced such a large difference. One possible reason is that they noticed the relatively small duration difference which native speakers produce in this pair and by analogy with other pairs of vowels came to regard this pair as another example of a long-short vowel pair. Another possibility is that they identified English /ɛ/ with Arabic short /i/ which is relatively close to /ɛ/ in terms of spectral properties. In fact, an examination of Figure 3-4 shows that

on the average the AR group did not produce a very large difference in quality between /ɪ/ and /ɛ/.

The data reported here confirm Mitleb's (1981) and Flege & Port's (1981) observation that Arabic speakers produce less of a duration difference between English vowels before voiced and voiceless stops than do native English speakers. It is generally believed that this pattern is the result of interference from Arabic. Specifically, it appears that vowel duration differences play only a minor role, if any role at all, in distinguishing post–vocalic consonant pairs in Arabic. It is apparently for this reason that Arabic speakers of English, even after years of experience, do not fully learn the importance of distinguishing post–vocalic voiced and voiceless consonants with vowel duration. In fact, Flege, Munro, & Skelton (under review) cite a number of studies in which native speakers of a variety of languages, including Arabic, French, Dutch, Finnish, and Japanese exhibited smaller voicing-conditioned duration differences than native speakers in their English vowel productions. In these studies, subjects with greater experience with English generally did not produce a more native-like duration difference.

One of the proposals made in Chapter 2 was that native speakers of Arabic may show an overall greater attentiveness to duration in vowel perception than native speakers of English. The findings of the production study neither strongly support nor strongly refute this proposal. On the one hand, if Arabic speakers are especially attentive to temporal properties of vowels, one might expect them to be highly attuned to *any* vowel duration differences. The fact that they exaggerate duration differences in English tense-lax pairs suggests that they mistakenly hear such pairs as long-short pairs, perhaps because they "match" them to Arabic long-short pairs. However, one might also expect them to readily notice that pairs such as *beat–bead* differ in vowel duration, and to produce such a pair

with a noticeable vowel duration difference. Yet this is not the case. In general the Arabic speakers here failed to produce a native-like pattern of duration differences in such pairs. While this may mean that they are not, in fact, especially aware of duration differences in vowels before voiced and voiceless stops, it is also possible that the Arabic speakers' use of phonemic vowel duration contrasts in their native language may interfere with the use of duration differences in post-vocalic stop production (see Flege, 1988), because a change in vowel duration is normally perceived as a change in vowel identity. In particular, lengthening the /ɪ/ of *bid* may result in a production which sounds too much like *bead* to an Arabic speaker. In fact, other researchers have reported that speakers of quantity languages often have difficulty learning to produce vowel duration differences before voiced and voiceless stops in English (e.g. Wiik, 1965).

***Vowel Quality***

The spectral data presented here indicate differences of some sort between the two groups on every vowel pair. To summarize briefly, the major differences observed between the two groups were in formant movement in all the front vowels, as well as in three of the five back vowels (/u/, /ʊ/, and /o/); in F1a values in /ɛ/, /ɑ/, and /ʌ/; and in F2a values in all the back vowels except /ɑ/.

While the extent to which the vowels produced by the Arabic speakers differed from the corresponding English vowels varied both from vowel to vowel and from subject to subject, it is rather striking that on the basis of the mean values from the AR group not a single vowel fell within the English norms on all four of the spectral parameters examined here, and all but two differed on at least two parameters. Even when individual subjects' data are considered, the majority of subjects show notable differences from the English mean values on one or more spectral parameters. These data provide rather convincing evidence that the effects of L1 experience on the

vowel production of adult L2 learners are very powerful and may be observable in some way in virtually every vowel from the L2 system. They also cast considerable doubt on the proposal that it is possible at least under laboratory conditions, and without special training, for a group of Arabic-speaking learners as a whole to show truly native-like production of *any* vowel from English, even after several years of experience. This, of course, does not mean that every individual vowel produced by an Arabic or other L2 learner of English is deviant in some way. Some productions may not show measurable differences from English norms, and some speakers may succeed in producing native-like vowels even though most do not. Moreover, even where measurable differences do exist, it is not necessarily the case that the vowel in question will be perceived by native listeners as accented. The question of how acoustic differences are related to native speakers' perceptions of accentedness will be addressed in Chapter 4.

The fact that differences in formant movement were observed in eight of the ten vowels examined here indicates that the Arabic speakers in general had not learned to produce English vowels with native-like patterns of diphthongization, a property of vowels which is usually overlooked in studies of English vowels. In most instances, the productions of the two groups differed because the AR vowels showed less movement than the EN vowels, although in a few instances the direction of movement was different. However, in no case did the AR speakers produce significantly *more* formant movement than the EN group. Although the AR group did show greater F1 and F2 movement in /e/ and /o/, traditionally considered diphthongs, than they did in other vowels, they still showed significantly less movement in these vowels than the EN speakers.

Descriptions of Arabic vowel articulation (eg., Mitchell, 1990) indicate that Arabic vowels, regardless of dialect, do not show as

much diphthongization as English vowels. The differences observed here, then, probably reflect an influence of L1 on L2 production. If speakers of Arabic were to substitute Arabic /i/ for English /i/, for instance, their production would involve less formant movement than would typically be found in the productions of native English speakers. Another possibility is that Arabic speakers are simply not very sensitive to formant movement from a perceptual perspective. When they learn a new vowel such as /ɛ/, they may fail to accurately note the formant movement patterns, and so may not produce them correctly.

The results presented here do not appear to support the distinction between "new" and "similar" phones proposed by Flege (1987b; 1988). Although most subjects failed to produce the similar vowels studied here (e.g. /i/, /u/) with native-like spectral properties as correctly predicted by the model, in the case of the vowel /ɛ/, which does not occur in some dialects and may therefore be considered "new," there was no evidence that the subjects from those dialect areas did better[2]. It might be proposed that in the case of several of the English vowels the Arabic speakers substituted native Arabic vowels, perhaps with slight modifications. While this is a plausible account, spectral data from all the dialects of Arabic represented here are not available, so the hypothesis cannot be readily tested.

***Relating Perception and Production***

The comparison between the perceptual data from Experiment I and the temporal and spectral measurements reported here was an exploratory analysis involving a small number of data points, and must therefore be viewed as preliminary work. It did indicate, nonetheless, that there is a relationship between the two data sets.

---

[2] It might also be noted that in Experiment IV (Chapter 4) the AR speakers' productions of this vowel were rated lower than those of any other vowel.

Two dimensions - F1 movement in /i/[3] and F2 frequency in /ɪ/ - were found to be especially important. The AR speakers who produced vowels which differed a great deal from those of native speakers with respect to these two dimensions were the same subjects who perceived the synthetic /i/-/ɪ/ continuum much less in terms of spectral properties than did the native speakers. This finding suggests that native-like perception of the /i/-/ɪ/ contrast may be a prerequisite for native-like production of the same contrast.[4] However, with respect to the F1a frequency of the vowel /i/, the observed correlation was in the direction opposite the one expected, a finding which appears to undermine this hypothesis. Specifically, the subjects who performed best in the perceptual task in Experiment I tended to show the largest differences from the English mean value of F1 in /i/. The reason for the reversal is unclear. One possibility is that the vowels with low values of F1 were not really "accented" vowels. In fact, the results of the regression analyses reported in Chapter 4 do indicate a small negative correlation between the accentedness ratings of native English judges and the F1 frequencies of the same vowel tokens examined here, when the effects of other predictor variables are partialled out. Furthermore, of the 12 subjects who exhibited relatively low (native-like) relative sensitivity ratios (< 0.075) and whose F1 values differed from the English mean by more than 20 Hz (recall that absolute values were used in the assessment of differences), 10 produced F1 values *lower* than the English mean. Taken together, these facts may indicate that the comparison value of F1 used here was too high. While the degree of change in the F1 of /i/ and the frequency of F2 in /ɪ/ may be

---

[3] Although there was no significant movement in F1 of /i/ in the EN group, several of the AR speakers showed considerable inappropriate movement on this dimension.

[4] The proposal that native-like perception is necessary for native-like production actually contradicts arguments raised by Sheldon & Strange (1982) and Gass (1984). For an extended discussion of this issue, see Chapter 5.

important determinants of accentedness in these vowels, the low value of F1a in /i/ may not result in any increase in accentedness. As long as an appropriate value of F2 was maintained, a low F1 would not result in any confusion of /i/ with another vowel, since there is no front vowel with a lower F1 than /i/. Therefore, it is possible, that the Arabic speakers who produced /i/ with a relatively low F1 were in fact producing native-like vowels.

## Tables and Figures for Chapter 3

**Table 3-1: Summary of Vowel Inventories by Dialect**

| Region | Phonemic Inventory | | Reference |
|---|---|---|---|
| | Long | Short | |
| Jordanian | i, ɑ, u | i, ɑ, u | Mitleb (1981) |
| Gulf | i, e, ɑ, u, o | i, ɑ, u | Holes (1990) |
| Egyptian | i, e, ɑ, u, o | i, ɑ, u | Gary & Gamal-Eldin (1984) |
| Syrian | i, e, ɑ, u, o | i, e, ɑ, u, o, ə | Cowell (1964) |
| Sudanese | i, e, ɑ, u, o | i, e, ɑ, u, o, ə | Trimingham (1959) |

**Table 3-2: Mean Durations by Vowel (ms)**

| GP | | /i/ | /ɪ/ | /e/ | /ɛ/ | /æ/ | /u/ | /ʊ/ | /o/ | /ɑ/ | /ʌ/ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AR | $\bar{x}$ | 205.8 | 128.7 | 225.0 | 138.6 | 232.7 | 221.6 | 154.5 | 230.3 | 158.8 | 145.5 |
| | SD | 29 | 28 | 30 | 22 | 40 | 43 | 37 | 32 | 36 | 35 |
| EN | $\bar{x}$ | 234.5 | 202.3 | 257.5 | 223.0 | 273.5 | 232.3 | 198.9 | 249.0 | 260.0 | 216.5 |
| | SD | 58 | 54 | 62 | 52 | 61 | 66 | 55 | 67 | 65 | 62 |

**Table 3-3: Mean Durations by Final Consonant (ms)**

| GP | | /t/ | /d/ |
|---|---|---|---|
| AR | $\bar{x}$ | 179.6 | 188.7 |
| | SD | 21 | 21 |
| EN | $\bar{x}$ | 192.2 | 277.3 |
| | SD | 33 | 42 |

**Table 3-4: Mean Formant Frequencies (Hz)**

**F1a Values**

| GP | | /i/ | /ɪ/ | /e/ | /ɛ/ | /æ/ | /u/ | /ʊ/ | /o/ | /ɑ/ | /ʌ/ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AR | $\bar{x}$ | 319 | 449 | 462 | 482 | 608 | 356 | 431 | 480 | 573 | 598 |
| | SD | 35 | 46 | 40 | 39 | 43 | 33 | 44 | 47 | 42 | 40 |
| EN | $\bar{x}$ | 322 | 436 | 471 | 521 | 634 | 370 | 436 | 505 | 669 | 567 |
| | SD | 24 | 26 | 32 | 31 | 49 | 30 | 35 | 19 | 34 | 28 |

**F2a Values**

| AR | $\bar{x}$ | 2150 | 1852 | 1940 | 1819 | 1643 | 917 | 1000 | 965 | 1075 | 1158 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | SD | 151 | 111 | 96 | 100 | 115 | 179 | 135 | 143 | 99 | 88 |
| EN | $\bar{x}$ | 2176 | 1837 | 1901 | 1769 | 1716 | 1370 | 1130 | 1192 | 1064 | 1263 |
| | SD | 117 | 96 | 105 | 61 | 74 | 185 | 94 | 149 | 48 | 96 |

**Table 3-5: Mean change in F1 and F2 (Hz)**

**ΔF1**

| GP | | /i/ | /ɪ/ | /e/ | /ɛ/ | /æ/ |
|---|---|---|---|---|---|---|
| AR | $\bar{x}$ | 5.6 | 3.4 | –45.0** | 7.5 | 18.6** |
| | SD | 17 | 25 | 45 | 21 | 17 |
| EN | $\bar{x}$ | –9.6 | 56.9** | –89.3** | 42.2** | 43.7** |
| | SD | 17 | 27 | 20 | 28 | 32 |

| GP | | /u/ | /ʊ/ | /o/ | /ɑ/ | /ʌ/ |
|---|---|---|---|---|---|---|
| AR | $\bar{x}$ | 6.8 | 10.2* | –31.9** | 8.6* | –3.1 |
| | SD | 17 | 18 | 39 | 18 | 23 |
| EN | $\bar{x}$ | –27.2** | 40.9** | –85.9** | 17.5* | 11.9 |
| | SD | 23 | 32 | 22 | 25 | 21 |

**ΔF2**

| GP | | /i/ | /ɪ/ | /e/ | /ɛ/ | /æ/ |
|---|---|---|---|---|---|---|
| AR | $\bar{x}$ | 49.4** | 13.5 | 136.0** | –10.1 | –29.5** |
| | SD | 60 | 56 | 128 | 34 | 46 |
| EN | $\bar{x}$ | 58.0** | –168.8** | 227.8** | –114.8** | –93.8** |
| | SD | 62 | 88 | 86 | 104 | 90 |

| GP | | /u/ | /ʊ/ | /o/ | /ɑ/ | /ʌ/ |
|---|---|---|---|---|---|---|
| AR | $\bar{x}$ | 65.0** | 129.9** | –3.0 | 125.0** | 116.1** |
| | SD | 82 | 87 | 95 | 75 | 81 |
| EN | $\bar{x}$ | –53.8 | 205.4** | –89.7** | 94.1** | 103.2** |
| | SD | 105 | 88 | 70 | 38 | 45 |

**$p < 0.01$; *$p < 0.05$ (two-tailed)

**Table 3-6: Comparison of Log-transformed Formant Values from the Two Groups (two-tailed *t* - values)**

| Vowel | F1a | F1a–F1b | F2a | F2–F2b |
|---|---|---|---|---|
| /i/ | –0.336 | _2.498* | –0.571 | 0.396 |
| /ɪ/ | 0.804 | 5.727*** | 0.367 | –7.149*** |
| /e/ | –0.656 | –3.295** | 1.100 | 2.405* |
| /ɛ/ | –2.907** | 3.619** | 1.503 | –4.248*** |
| /æ/ | –1.552 | 2.692* | –2.001 | –2.693* |
| /u/ | –1.228 | –4.628*** | –6.746*** | –3.536** |
| /ʊ/ | –0.356 | 3.455** | –2.955** | 1.469 |
| /o/ | –1.856 | –4.234*** | –4.203*** | –2.670* |
| /ɑ/ | –6.567*** | 0.896 | 0.246 | –1.251 |
| /ʌ/ | 2.415* | 1.892 | –3.158** | –0.793 |

***$p < 0.001$; **$p < 0.01$;*$p < 0.05$

**Table 3-7: Comparison of Individual Duration Data and Perception Scores from Experiment I**

| S # | Reg. | /i/ | /ɪ/ | i/ɪ ratio | /e/ | /ɛ/ | e/ɛ ratio | /u/ | /ʊ/ | u/ʊ ratio | RS** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | JOR | 216.4 | 108.5 | 2.0 | 228.2 | 123.1 | 1.9 | 167.2 | 134.6 | 1.2* | 0.2244 |
| 2 | KUW | 228.4 | 171.6 | 1.3* | 257.3 | 141.4 | 1.8 | 218.0 | 136.8 | 1.6 | 0.0320 |
| 3 | JOR | 196.9 | 117.8 | 1.7 | 186.9 | 142.7 | 1.3 | 241.4 | 134.3 | 1.8 | 0.0222 |
| 4 | KUW | 179.6 | 146.0 | 1.2* | 199.8 | 143.0 | 1.4 | 194.1 | 207.6 | 0.9 | 0.0256 |
| 5 | KUW | 211.8 | 91.2 | 2.3 | 223.5 | 147.8 | 1.5 | 213.0 | 132.6 | 1.6 | 0.0040 |
| 6 | JOR | 212.5 | 160.6 | 1.3* | 250.3 | 171.9 | 1.5 | 278.6 | 199.5 | 1.4 | 0.0420 |
| 7 | SYR | 190.7 | 107.4 | 1.8 | 203.9 | 102.4 | 2.0 | 195.0 | 123.8 | 1.6 | 0.0086 |
| 8 | KUW | 166.6 | 98.6 | 1.7 | 189.6 | 111.3 | 1.7 | 178.9 | 136.7 | 1.3* | 0.0290 |
| 9 | SAU | 192.1 | 129.5 | 1.5 | 295.8 | 134.9 | 2.2 | 196.6 | 140.2 | 1.4 | 0.2737 |
| 10 | JOR | 220.9 | 102.2 | 2.2 | 253.2 | 119.7 | 2.1 | 274.2 | 130.1 | 2.1 | 0.1964 |
| 11 | SAU | 207.3 | 192.8 | 1.1* | 232.9 | 162.7 | 1.4 | 304.9 | 180.7 | 1.7 | 0.0084 |
| 13 | KUW | 221.0 | 169.3 | 1.3* | 261.9 | 126.7 | 2.1 | 197.0 | 136.4 | 1.4 | 0.0169 |
| 14 | SUD | 255.9 | 75.2 | 3.4 | 220.0 | 104.8 | 2.1 | 221.7 | 99.5 | 2.2 | 0.4188 |
| 15 | SUD | 181.7 | 127.3 | 1.4 | 216.9 | 148.2 | 1.5 | 207.3 | 198.9 | 1.0* | 0.2116 |
| 16 | SUD | 192.3 | 98.4 | 2.0 | 160.2 | 135.3 | 1.2* | 135.1 | 117.6 | 1.1* | 0.2480 |
| 17 | PAL | 206.4 | 154.3 | 1.3* | 227.3 | 133.8 | 1.7 | 231.5 | 142.7 | 1.6 | 0.0121 |
| 18 | EGY | 197.9 | 132.4 | 1.5 | 230.7 | 133.6 | 1.7 | 217.3 | 133.4 | 1.6 | 0.0147 |
| 19 | SYR | 196.8 | 122.4 | 1.6 | 249.2 | 165.6 | 1.5 | 252.9 | 175.8 | 1.4 | 0.0606 |
| 20 | JOR | 181.8 | 118.9 | 1.5 | 181.3 | 108.5 | 1.7 | 175.8 | 128.6 | 1.4 | 0.0264 |
| 21 | KUW | 180.9 | 126.9 | 1.4* | 196.5 | 131.2 | 1.5 | 197.1 | 178.4 | 1.1* | 0.1203 |
| 22 | SAU | 172.5 | 97.7 | 1.8 | 196.0 | 89.0 | 2.2 | 157.5 | 105.4 | 1.5 | 0.0089 |
| 23 | SUD | 177.2 | 116.7 | 1.5 | 184.1 | 135.6 | 1.4 | 189.7 | 118.0 | 1.6 | 0.2882 |
| 24 | PAL | 227.9 | 190.0 | 1.2* | 224.1 | 188.0 | 1.2* | 204.7 | 195.3 | 1.1* | 0.0268 |

*Denotes ratio within one standard deviation of English mean.
**Relative sensitivity to duration in Experiment I (see text).

**Table 3-8: Individual Differences from English Means on F1 and F2 of /i/ and /I/ (Hz)**

| Subj # | Region | F1 /i/ | F2 /i/ | F1 /I/ | F2 /I/ |
|---|---|---|---|---|---|
| 1 | JOR | 22* | 162 | 42 | 7* |
| 2 | KUW | 14* | 42* | 8* | 87* |
| 3 | JOR | 46 | 155 | 25* | 18* |
| 4 | KUW | 48 | 233 | 51 | 136 |
| 5 | KUW | 50 | 198 | 49 | 57* |
| 6 | JOR | 74 | 265 | 137 | 61* |
| 7 | SYR | 5* | 87* | 51 | 1* |
| 8 | KUW | 29 | 198 | 9* | 128 |
| 9 | SAU | 3* | 129 | 8* | 16* |
| 11 | SAU | 38 | 95* | 59 | 10* |
| 13 | KUW | 21* | 61* | 68 | 42* |
| 14 | SUD | 22* | 188 | 18* | 310 |
| 15 | SUD | 31 | 35* | 44 | 207 |
| 16 | SUD | 14* | 95* | 25* | 153 |
| 17 | PAL | 46 | 181 | 9* | 85* |
| 19 | SYR | 29 | 16* | 44 | 104 |
| 20 | JOR | 29 | 188 | 59 | 104 |
| 21 | KUW | 14* | 95* | 69 | 138 |
| 22 | SAU | 21* | 85* | 17* | 53* |
| 23 | SUD | 3* | 68* | 78 | 242 |
| 24 | PAL | 46 | 77* | 35 | 10* |

*Denotes difference within one standard deviation of English mean.

**Table 3-9: Individual Spectral Differences on Three Vowels**

| Subj # | Region | Movement in F1 of /e/ | F1 Diff.* /ɛ/ | F1 Diff.* /ɑ/ |
|---|---|---|---|---|
| 1 | JOR | –78 | –53 | –157 |
| 3 | JOR | 56 | –57 | –21 |
| 6 | JOR | –25 | –2 | –72 |
| 20 | JOR | –56 | –10 | –120 |
| 2 | KUW | –73 | –45 | –115 |
| 4 | KUW | –69 | –10 | –107 |
| 5 | KUW | –55 | –27 | –70 |
| 8 | KUW | –64 | –27 | –102 |
| 13 | KUW | –86 | –27 | –21 |
| 21 | KUW | –9 | –139 | –124 |
| 17 | PAL | –18 | –53 | –115 |
| 24 | PAL | –65 | –100 | –111 |
| 9 | SAU | –39 | –36 | –90 |
| 11 | SAU | –98 | –36 | –60 |
| 22 | SAU | –91 | –27 | –51 |
| 14 | SUD | –107 | 63 | –34 |
| 15 | SUD | 26 | –49 | –163 |
| 16 | SUD | –95 | –49 | –133 |
| 23 | SUD | –5 | –23 | –85 |
| 7 | SYR | 8 | –32 | –150 |
| 19 | SYR | –4 | –83 | –111 |

* Compared to mean value for EN Group

**Table 3-10: Correlations between Perceptual Data (Experiment I) and Spectral Properties**

| Vowel | Acoustic Variable | *r* |
|---|---|---|
| /i/ | F1 | –0.467* |
| | ΔF1 | 0.430* |
| | F2 | –0.069 |
| | ΔF2 | 0.257 |
| /ɪ/ | F1 | –0.128 |
| | ΔF1 | –0.001 |
| | F2 | 0.638** |
| | ΔF2 | 0.322 |

** $p<0.01$; *$p<0.05$ (two-tailed)

**Figure 3-1: Durations of Front Vowels**

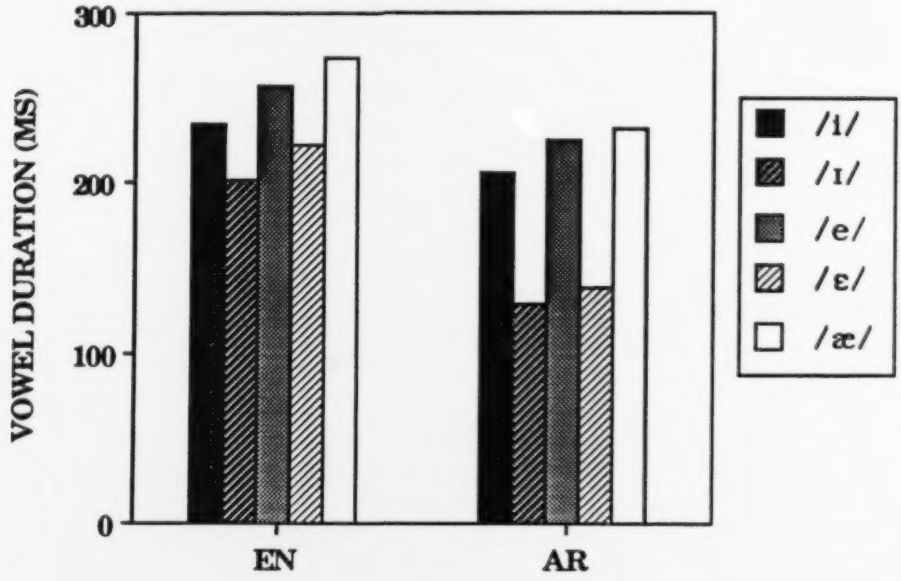



**Figure 3-2: Durations of Back Vowels**

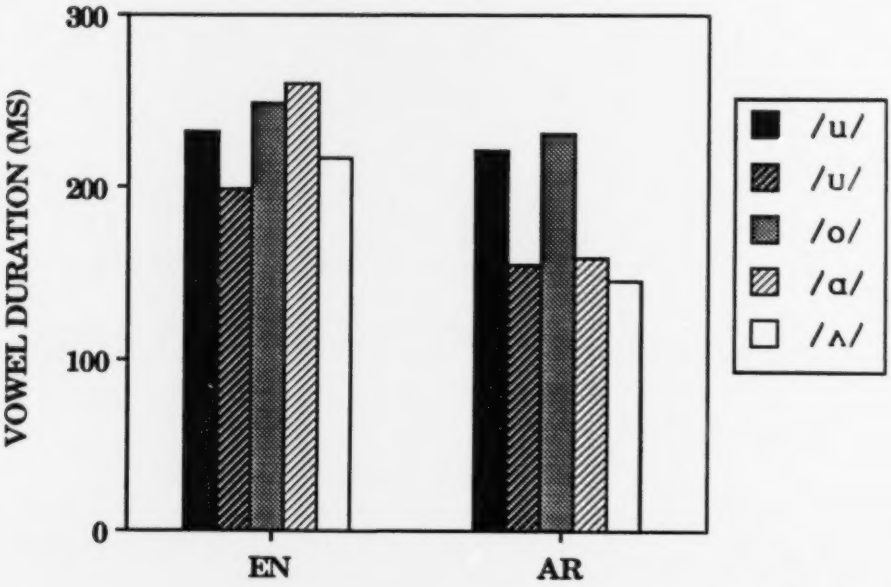


**Figure 3-3 : Vowel Duration by Final Consonant**

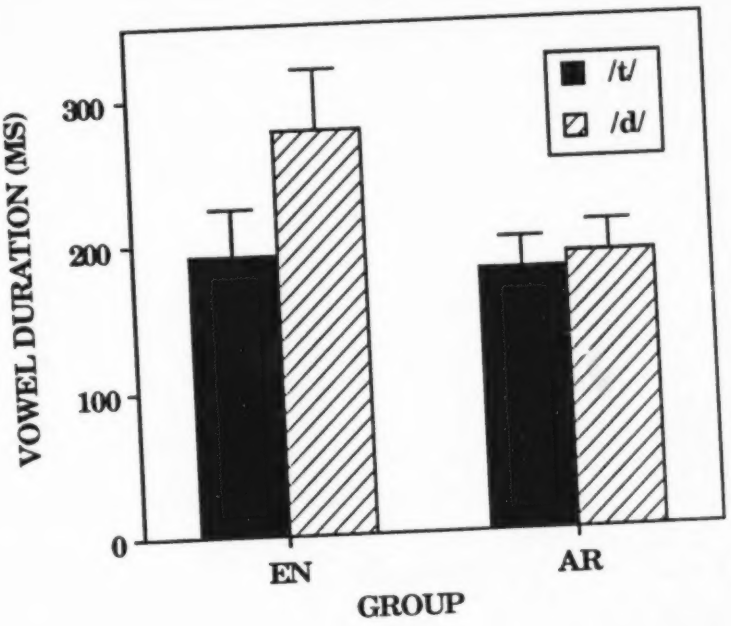



**Figure 3-4: Mean Formant Values (Measurement A)**

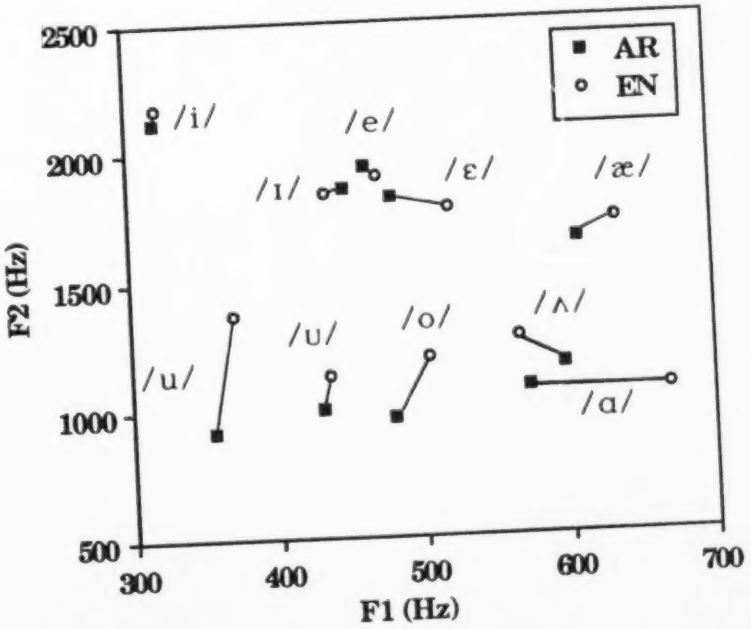


**Figure 3-5: Change in F1**

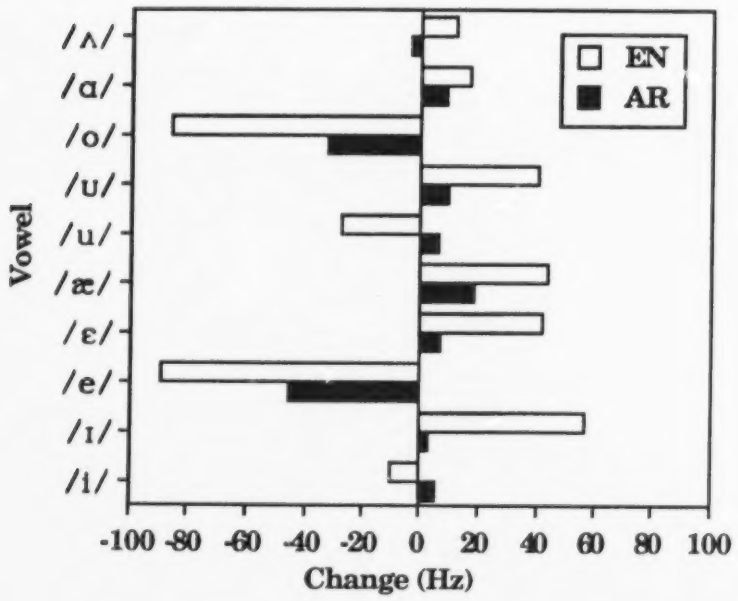
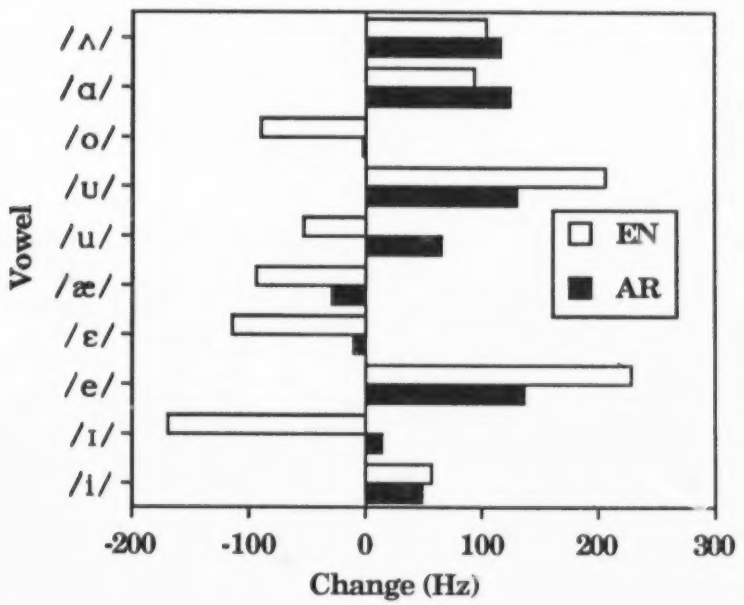


**Figure 3-6: Change in F2**



**References**

Al-Ani, S. (1970). *Arabic phonology* . The Hague: Mouton.

Andruski, J., & Nearey, T. (in press). On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables. *Journal of the Acoustical Society of America*.

Bohn, O-S., & Flege, J. (1990). Interlingual identification and the role of foreign language experience in L2 vowel perception. *Applied Psycholinguistics* **11**: 303-328.

Caramazza, A., Yeni–Komshian, G., Zurif, E., & Carbone, E. (1973). The acquisition of a new phonological contrast: the case of stop consonants in French-English bilinguals. *Journal of the Acoustical Society of America* **54**: 421-428.

Cowell, M. (1964) *A Reference Grammar of Syrian Arabic* . Washington: Georgetown University Press.

Crystal, T., & House, A. (1988). The duration of American-English vowels: an overview. *Journal of Phonetics* **16**: 263-284.

Disner, S. (1986). On describing vowel quality. In J. Ohala and J. Jaeger (Eds.), *Experimental Phonology* (pp 69-79). Orlando: Academic Press.

Flege, J. (1980). Phonetic approximation in second language acquisition. *Language Learning* **30**: 117-134.

Flege, J. (1987a). The instrumental study of L2 speech production: some methodological considerations. *Language Learning* **37**: 285-296.

Flege, J. (1987b). The production of "new" and "similar" phones in a foreign language: evidence for the effect of equivalence classification. *Journal of Phonetics* **15**: 47-65.

Flege, J. (1988). The production and perception of foreign languages. In H. Winitz (Ed.), *Human communication and its disorders* (pp. 224-401). Norwood, NJ: Ablex.

Flege, J., & Hillenbrand J. (1984). Limits on phonetic accuracy in foreign language speech production. *Journal of the Acoustical Society of America* **76**: 708-721.

Flege, J., Munro, M., & Skelton, L. (under review). Production of the word-final English /t/-/d/ contrast by native speakers of Mandarin and Spanish. *Journal of the Acoustical Society of America* .

Flege, J., & Port, R. (1981). Cross-language phonetic interference: Arabic to English. *Language and Speech* **24**: 125-146.

Gary, J., & Gamal–Eldin, S. (1984). *Cairene Egyptian Colloquial Arabic*. London: Croom Helm.

Gass, S. (1984). Development of speech perception and speech production abilities in adult second language learners. *Applied Psycholinguistics* **5**: 51-74.

Hindle, D. (1978). Approaches to vowel normalization in the study of natural speech. In D. Sankoff (Ed.). *Linguistic variation: models and methods* (pp. 161-171). New York: Academic Press.

Holden, K., & Nearey, T. (1986). A preliminary report on three Russian dialects: Vowel perception and production. *Russian Language Journal* **40**: 3-21.

Holes, C. (1990). *Gulf Arabic* . New York: Routledge.

Jamieson, D., Ramji, K., Nearey, T., & Baxter, T. (1990). *Canadian Speech Research Environment: User's Manual*. London, ON: Speech Communication Laboratory, University of Western Ontario.

Jonasson & McAllister (1972). Foreign accent and timing: an instrumental phonetic study. *Papers from the Institute of Linguistics* (PILUS), Publication 14, pp. 11-40. Stockholm: University of Stockholm.

Mitchell, T. (1990). *Pronouncing Arabic I*. Oxford: Clarendon Press.

Mitleb, F. (1981). Timing of English vowels spoken with an Arabic accent. *Research in Phonetics*, Report No. 2, pp. 193-226. Bloomington:Indiana University, Department of Linguistics.

Nearey, T. (1977). *Phonetic feature systems for vowels.* Bloomington, IN: Indiana University Linguistics Club.

Nearey, T., & Assmann, P. (1986). Modelling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America* **80**: 1297-1308.

Port, R., & Mitleb, F. (1983). Segmental features and implementation in acquisition of English by Arabic speakers. *Journal of Phonetics* **11**: 219-229.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics* **3**: 243-261.

Trimingham, J. (1959). *Sudan Colloquial Arabic.* London: Oxford University Press.

Umeda, N. (1975). Vowel duration in American English. *Journal of the Acoustical Society of America* **58**:434-445.

Williams, L. (1980). Phonetic variation as a function of second-language learning. In G. Yeni-Komshian, J. Kavanagh, & C. Ferguson (Eds.), *Child phonology, Vol. 2: Perception* (pp. 185-215). New York: Academic Press.

Wiik, K. (1965). *Finnish and English Vowels.* Turku: Turun Yliopiston Julkaisuja Annales Universitatis Turkuensis.

Winer, B. (1971). *Statistical principles in experimental design* . (2d ed.). New York: McGraw-Hill.

# CHAPTER 4

# EXPERIMENT IV: ACCENTEDNESS JUDGMENTS AND ACOUSTIC PROPERTIES OF FOREIGN-ACCENTED VOWELS

## Introduction

In Experiment III some of the differences between the vowel productions of native and non–native speakers of English were quantified in terms of spectral and temporal properties. Even though the majority of the non-native subjects in the sample had been using English for five or more years, their productions differed significantly from those of the native speakers in some respect on virtually every vowel examined. Moreover, an examination of the performance of individual L2 subjects on selected vowels showed that few subjects could be viewed as producing these vowels in a native-like way.

While the measurement data collected in Experiment III reveal a number of between-group differences in English vowel production, they say nothing about how those differences are actually perceived by native listeners. In particular, they do not indicate which differences contribute to the perception of a foreign accent or how large those differences must be before they are noticeable to native speakers. In order to answer such questions it is necessary to examine native speakers' evaluations of accentedness in English vowel productions and attempt to relate those evaluations to acoustic properties.

### *Native Speakers' Assessments of Non-native Productions*

Several approaches have been used in studies of how native speakers perceive the utterances of L2 learners. One technique is simply to have native-speaking listeners identify single sounds or words produced by non-native speakers. The identification rates reflect whether or not the speakers produce a particular sound or

sounds sufficiently well to be well identified. Sheldon & Strange (1982), for instance, used correct identification rates of Japanese speakers' productions of /r/ and /l/ obtained from native English listeners to assess how well the Japanese speakers produced this contrast. Flege & Hillenbrand (1984) had native speakers of French identify tokens of the French words /tu/ and /ty/ produced by native English speakers of French. They observed that more experienced speakers generally produced a more effective contrast between the two vowels. They also observed a fairly high correlation between identification scores and F2 frequencies. Flege, Munro, & Skelton (under review) had native English speakers identify English voiced and voiceless final alveolar stops produced by speakers of Mandarin and Spanish. A variety of predictor variables, including vowel duration, amount of closure voicing, and F1 offset frequency were then regressed on the identification scores in an effort to determine which acoustic properties were correlated with identifiability.

In a different approach, used by Flege (1984), listeners judged whether speakers of utterances of various durations were native or non-native speakers of English. The listeners (both phonetically trained and untrained) were generally successful at the task, whether they were presented with three-word phrases, CV words, or even short (30 ms) /t/ bursts produced by native French speakers of English and native English speakers. His finding that naive native speakers were able to notice accentedness even in very short segments of speech, such as bursts, indicates a high degree of sensitivity to divergences from native speaker norms, at least for consonants. A similar finding was reported by Munro (1987), who observed that native English speakers were sensitive to small differences in voice onset time in natural productions of unaspirated French /p/ and English /b/, even though these sounds are very similar.

In Experiment IV a group of linguistically-trained listeners directly rated English vowels produced by native Arabic speakers on a scale of "accentedness." In several other studies which have used this approach, various scales and lengths of utterances have been used, and both phonetically-trained and naive listeners have served as judges. It has been shown that accentedness ratings from linguistically-trained and naive listeners are generally consistent with one another (Brennan & Brennan, 1981; Cunningham-Andersson & Engstrand, 1989). Often, researchers have used this method to assess the role of various predictor variables on general success in second language pronunciation. Suter (1976) and Purcell & Suter (1980), for instance, found their subjects' first language, aptitude for oral mimicry, and concern for pronunciation accuracy to be significant predictors when they regressed data describing a variety of speaker characteristics on scores assigned by trained judges using a six-point scale. The ratings were based on a two-minute speech sample, obtained from the talkers by having them speak on a topic without formal preparation. Tahta, Wood, & Loewenthal (1981) found the age at which English was learned to be highly correlated with pronunciation scores assigned to 109 English L2 speakers who read passages from a text. Flege (1988) used a much larger rating scale. His naive listeners rated short sentences read from a list by positioning a lever on a response box with 8-bit resolution, giving an effective rating scale of 1 to 256. He observed that the utterances of English learners were generally rated lower than those of native speakers, even when the L2 subjects had learned English in childhood. He also observed no difference in pronunciation scores between subjects who had lived in the United States for one year and those who had lived there for 5 years.

***Relating Accentedness Ratings to Acoustic Properties***

In the present study, certain acoustic properties of accented English vowels will be regressed on accentedness ratings assigned by a panel of judges. The purpose of this research is to determine which properties are associated with accentedness in a set of English front vowels and how well accentedness ratings can be predicted from them. Research by Greiser & Kuhl (1989) supports the notion that listeners have access to prototypes for vowel categories. Their data indicate that even six-month-old infants show better generalization of 'good' exemplars of a vowel (as defined by adult ratings) than of 'poor' exemplars. When native speakers are asked to rate accentedness of individual speech sounds produced by non-natives, they may make reference to such prototypes. Vowel tokens which differ from a prototype along such dimensions as F1 and F2 frequencies, degree of formant movement, and duration, may be rated as having varying degrees of accentedness, depending on how much they differ from the prototype along these dimensions.

In Experiment IV, accentedness ratings of five vowels (/i/, /ɪ/, /e/, /ɛ/, and /æ/) produced by the non-native speakers in Experiment III will be collected. If a "prototype account" of accentedness ratings is correct, it might be possible to use a multiple regression analysis to observe a relationship between some of the acoustic properties of the tokens and the judges' ratings. For instance, tokens of the vowel /i/ having some values of F1 and F2 should be rated as more accented than others because they differ more along those dimensions from a prototypical /i/.

One difficulty peculiar to vowel description in a study such as this is the question of how best to capture the important properties of vowel productions so that meaningful between-group comparisons can be made. Many cross-language and foreign accent studies have attempted to characterize vowel productions, following Peterson & Barney (1952), with single measurements, usually taken at vowel

midpoints, of F1 and F2 (Disner, 1986; Flege, 1987). A drawback to this method is that vowels are not in fact "steady-states," and different languages and accents may exhibit important differences in formant movement. For instance it is known that Arabic vowels do not, as a rule, show as much formant movement as English vowels (Mitchell, 1990), and as the results of Experiment III (Chapter 3) indicate, speakers of Arabic tend to transfer this characteristic of Arabic vowels to their productions of most English vowels. Since a lack of appropriate formant movement may be a salient characteristic of accented vowels, it is highly desirable to incorporate some evaluation of formant movement into vowel descriptions. The experiment reported here will do so.

A further problem in this study is the question of how best to transform the acoustic predictor variables so that it is reasonable to expect correlations with the rating data. The simple use of raw formant values in a linear analysis is unlikely to be a successful approach. To see why this is so, consider Figure 4-1. This figure gives a highly idealized rating function (defined by a second order polynomial) which might approximate rating data on some vowel. Suppose that for this vowel only one predictor (F1 frequency for instance) is required to account for the ratings assigned. In this case, one particular value of F1 will yield the highest rating, while values above and below this ideal value will correspond to lower ratings. If this relationship is like the type of relationship that actually exists between formant frequencies (or durations) of vowel tokens and corresponding accentedness ratings, one cannot expect linear regression with raw formant frequencies to yield positive results.

One way to handle this problem is to choose a set of production data likely to be representative of prototypical vowel categories and to quantify the differences between the non-native data and this standard. In other words an assumption is made about what the

ideal value of some acoustic dimension (e.g., F1) is. The differences may then be used as predictor variables, with the assumption that large differences will be correlated with a high degree of perceived accentedness. In this study, the Arabic speakers' production data will be compared against corresponding data collected from native English speakers. It will be assumed, then, that the native data are fairly representative of prototypical vowels. There are obvious disadvantages to such an assumption. For instance, data from a relatively small sample of speakers may not actually be representative of the prototypes listeners have access to. However, it is a reasonable starting point for this type of research. If a relationship between the two data sets is observed, there will be good reason to attempt to refine the model to give a better account of the accentedness ratings.

## Methods

### *Materials*

The stimuli were 115 /bVt/ tokens which were recorded and analysed in Experiment III. The tokens used contained the vowels /i/, /ɪ/, /e/, /ɛ/ and /æ/, produced by 21 male speakers from the AR group and 2 male speakers from the EN group, for a total of 23 tokens containing each vowel. The two native English speakers were selected on the basis of F1 and F2 values which were close to the English means. One of the native English speakers was a native of Birmingham, Alabama; the other was from Georgia.

### *Procedure*

The /bVt/ tokens were evaluated by a panel of five native English, linguistically-trained judges, who rated each vowel on an accentedness scale. Tokens were blocked by vowel and presented randomly in five listening sessions of about seven minutes each. Five replications of each token were presented in each session, for a total

of 115 tokens per session. A total of 2875 judgments were collected, 25 on each token. In each session, the listeners were told in advance which vowel they were to evaluate. They were given no other information about the actual words they were hearing, but were told that the speakers were sometimes native and sometimes non-native speakers of English. Each stimulus was presented twice by an IBM AT microcomputer through a loud speaker in a sound–treated room. The judges assigned accentedness ratings by placing a cursor on a rating scale displayed on the computer screen. Although there were actually 101 points on the rating scale, only 3 labels were used: "0" at the far left, "100" at the far right, and "50" in the middle. The judges were instructed to attend to the vowel portion of the stimulus and to position the cursor at 100 if the vowel sounded like a perfectly natural, unaccented exemplar of the target vowel. They were to position the cursor at 0 if the vowel did not sound at all like the target vowel. Other positions on the scale could be used to indicate varying degrees of accentedness, with positions on the left indicating more accentedness than positions on the right. The listeners were advised that they did not have to use the entire scale in each session.

Of course, the 101 points on the scale probably represent a far greater number of "goodness distinctions" than the listeners could actually make. However, this number was convenient because, from the point of view of the subjects, it effectively permitted a continuous rating scale. Since only three anchors were given on the scale, the subjects were encouraged to relate their judgments visually to positions on the scale, rather than to attempt to assign numerical ratings to stimuli. This procedure is therefore very similar to the one used by Flege (1988) to obtain foreign accent ratings from naive native speakers.

**Results**

The ratings are organized by vowel and speaker in Table 4-1. Mean ratings by vowel were obtained by pooling the ratings over the non-native speakers only. The highest-rated vowel was /i/ (71), while the lowest-rated was /ɛ/ (39). Mean ratings for each speaker pooled over vowels are also given. The two native speakers received higher overall ratings than any other speakers, 92 and 86, while the Arabic speakers' ratings ranged from 40 to 78. The fact that the standard deviations of the ratings given to the native speakers (based on all 25 ratings assigned to each speaker) were smaller than those of the ratings given to most other subjects suggests relatively strong agreement amongst the judges that these speakers produced unaccented vowels, although it should be pointed out that variability is likely to be lower in high scores anyway because of a ceiling effect.

The ratings by speaker on each of the five vowels are also given in Table 4-1. In the case of /i/, six Arabic speakers received ratings equal to or higher than one of the native speakers. Two Arabic speakers received such scores on /e/. From the point of view of the judges, then, these eight tokens, and possibly a few others with relatively high ratings, were as good as native speakers' vowels. However, on the vowels /ɪ/, /ɛ/, and /æ/, every non-native production was rated lower than both native productions. Across vowels, on only 13 of 105 tokens (12%) did a non-native speaker receive a score greater than 80. In the vast majority of cases, then, it appears that the judges had little difficulty in distinguishing native from non-native productions.

In order to assess whether any of the judges differed radically from the others in the pattern of ratings assigned to the speakers, Pearson correlation coefficients were calculated on the ratings for all possible pairs of judges (Table 4-2). Since all these correlations are in the range of 0.75 to 0.83, and are therefore significant at well beyond

the 0.01 level (N=115), it appears that the judges were quite consistent with one another.

***Predictor Variables***

The acoustic measurements of the vowel tokens were made in Experiment III using LPC analysis and cursor measurements from a CRT display. F1 and F2 frequencies were measured to the nearest 1 Hz at two points, once at approximately 30% of the distance into the vowel and once at about 70%. These were labelled the 'a' and 'b' measurements respectively (see Chapter 3). Vowel duration was measured to the nearest 0.1 ms. The formant frequencies and durations were transformed as described below and used as predictor variables in a stepwise linear multiple regression analysis with the mean accentedness ratings as the dependent variable. It is reasonable to expect that the rating given to any particular token will depend on such factors as formant frequencies, formant movement, and possibly duration. In the present experiment, the properties of interest were the F1a and F2a values (here labelled *F1* and *F2*), F1 and F2 movement between the 'a' and 'b' measurements ($\Delta F1$ and $\Delta F2$) and vowel duration (*DUR*).

In the primary analysis undertaken here (hereafter referred to as Analysis I), the standard against which the vowel tokens were compared on any parameter was determined by taking the mean value of the corresponding measurements from the 12 male native speakers whose production data were examined in Experiment III. Six predictors were used. Five of the predictor values were computed from raw measurement data using the formulae in Table 4-3. Symbols subscripted with '$\bar{x}$' are mean values from the EN group. Values of the *F1* and *F2* predictor variables were calculated by subtracting the $F1a_{\bar{x}}$ and $F2a_{\bar{x}}$ values from the values of F1a and F2a on the 115 individual tokens. These values, then, represent an assessment of the "distance" between the individual speakers' tokens

and the English means. To compare the degree of change in formant frequencies (*ΔF1* and *ΔF2*), the F1a and F2a measurements on individual tokens were subtracted from the corresponding 'b' measurements, and the differences between the native mean values were subtracted from the results. Finally, the vowel duration comparisons were made by subtracting the durations of individual tokens from the native means. For each of these five predictor variables, squares of the resulting differences were used so that a relatively large value of any parameter indicates a large discrepancy between it and the native mean, regardless of direction.[1] The regression analysis is therefore a way of testing for hypothetical rating functions like the one shown in Figure 4-1.

It should be stressed that this analysis is exploratory since no fully developed theoretical model exists against which the results can be compared. Therefore, the formulae and transformations employed here were chosen simply because of their intuitive plausibility. In addition, the categorial variable *VOWL* was included in the analysis, to determine whether listeners showed a different pattern of results on different vowels.

### *Regression Analyses*

#### *Analysis I*

In the stepwise analysis with F-to-enter set to 4.00, four of the six predictor variables were sufficiently good predictors to be entered by the program in a total of four steps. They were, in the order in which they were entered, $F1^2$, $\Delta F1^2$, $\Delta F2^2$, and *VOWL*. At first it may seem surprising that the predictor with the highest partial correlation was $F1^2$, given that Experiment III revealed that the differences between the two groups on the F1a measurements of the

---

[1] In the case of *ΔF1* and *ΔF2* , if a token exhibits formant movement increasingly in the direction opposite to that of the EN mean value, the value of *ΔF1* or *ΔF2* will increase accordingly.

front vowels were mostly non-significant. It should be remembered, however, that even though, as a group, the AR subjects may not have differed significantly on the F1a values, individual productions differed to varying degrees from the native English mean on this dimension. The analysis yielded a moderate multiple $R$ of 0.670 which indicates that the model accounted for 43% of the variance (adjusted for degrees of freedom).

The fact that *VOWL* was a significant predictor variable suggests that the judges may have weighed the acoustic cues somewhat differently depending on vowel identity. To explore this issue further, independent stepwise analyses were carried out on each vowel with the five remaining predictor variables. The results are given in the first row of Table 4-4. In the case of /i/, none of the predictor variables could be entered into the analysis.[2] The only predictor which was entered in the analyses of the /ɪ/ , /ɛ/, and /æ/ ratings was $F1^2$, which yielded $R^2$ values of 28%, 30%, and 34% respectively. In the analysis of the /e/ ratings, a relatively high $R^2$ of 78% was calculated from two predictor variables: $\Delta F2^2$ and $DUR^2$.

In summary, these results suggest that formant movement figured strongly in the subjects' ratings of /e/, while accentedness in the other front vowels was correlated with the frequency of F1. As observed in Experiment III, the Arabic speakers, as a group, produced /ɛ/ with significantly different F1 values from those of the native speakers. In fact, all but two had F1 frequencies *below* the English mean value. While there was no significant difference between the two groups in the F1 frequency of /æ/, 17 of the 21 Arabic speakers produced this vowel with an F1 frequency below the English mean. It is not surprising that front vowels with relatively low F1 values should be rated as accented. In traditional terms, the effect of lowering F1 is to raise a vowel in the F1-F2 space. This may have

---

[2]The highest partial correlations were found with $\Delta F2$ (-0.389) and $\Delta F1$ (–0.340).

caused the listeners to hear the vowel as more ambiguous than others or even as belonging to a category "higher" than the one intended. In debriefing, in fact, some of the judges commented that a few of the /ɛ/ tokens sounded like /ɪ/.

In general, the method used here was moderately successful in accounting for the ratings assigned to the Arabic speakers' productions. It was clearly demonstrated that there was a relationship between the acoustic properties of the vowels and the judges' accentedness ratings. However, it is conceivable that better predictions of the rating data might be obtained if modifications were made to the prediction model. The assumption that the English speakers' production data were representative of prototypes used by listeners may, of course, have led to less satisfactory results than might have been obtained if other standard values had been used. Therefore another approach involving different assumptions about the data was tried.

*Analysis II*

In Analysis II, the squares of the differences between individual values and the native English means on the five acoustic variables were used just as in Analysis I. However, this time the raw differences were also entered into the analysis as additional predictors. This model takes into account the possibility that the native English mean values on the acoustic variables are not exactly equal to the ideal values of those variables, and, in effect, allows for a correction due to inaccurate estimates of the ideal values.

Tables 4-4 and 4-5 compare the results of the two analyses. Table 4-5 gives the relevant $R^2$ values for all vowels together and each vowel separately, while Table 4-6 indicates which predictors were selected by the stepwise procedure. As can be seen from the tables, Analysis II gave exactly the same results on the vowel /e/, since the same two predictors were chosen. It was slightly more successful in

predicting the judges' ratings of the other four vowels, however. In the case of /i/ two variables, *F2* and $\Delta F2^2$, were entered. For /ɪ/, $F1^2$ proved again to be the best predictor, but a slightly higher $R^2$ value was obtained because of the addition of $\Delta F2$. In Analysis II, *F1* (not $F1^{2)}$ proved to be the best predictor of accentedness in /ɛ/, and /æ/. The new model appeared to perform slightly better on /ɛ/ using only *F1* as a predictor and better on /æ/ with $\Delta F1$ included.

In Analysis II a linear relationship was observed between the raw values of the *F1* predictor variable and the ratings on /ɛ/ and /æ/. While one might be tempted to conclude that this provides evidence against a second order rating function such as that shown in Figure 4-1, a closer examination of the data indicates that almost all subjects produced low F1 values for these vowels (19 in the case of /ɛ/ and 17 for /æ/). If several subjects had produced F1 values much higher than the native English mean, their productions might have received lower ratings and a second order relationship might have been observed. Therefore, this finding is not evidence against the proposed rating function. It does illustrate, however, one of the difficulties inherent in the use of natural as opposed to synthetic stimuli: the experimenter has no control over the range of formant frequencies in the speakers' productions.

In the case of /i/, a linear relationship was observed between F2 frequency and the mean ratings. Since the correlation was positive, it appears that higher F2 frequencies were associated with higher ratings. This is not entirely surprising, since, if other factors remained constant, an increase in the F2 frequency of /i/ , even beyond some prototypical value, would not lead to any confusion with other vowels.

It is clear that many other models could be tested here, but it should be stressed that these analyses are exploratory, and much further data will have to be collected and analyzed before conclusions can be drawn about which approach is the best.

***Relating Accentedness Ratings to Other Data***

It might be expected that the subjects who perceived the English /i/–/ɪ/ continuum in Experiment I in the most native-like way might also be the subjects who scored the highest ratings on these vowels in Experiment IV. In order to test this hypothesis, the relative sensitivity ratios calculated in Experiment I were compared with the mean ratings on the vowels /i/ and /ɪ/ obtained in Experiment IV. Recall that large ratios indicate a high relative sensitivity to duration in the perception of a synthetic /i/–/ɪ/ continuum, and therefore, a relatively non-native pattern of results. Non-significant Pearson correlations of -0.157 and 0.191 respectively were obtained, indicating no relationship between the perceptual data and the ratings.

It might also be the case that amount of experience with English is correlated with accentedness. To test this proposal, Pearson correlations were calculated between the YOE (years of English), LOR (length of residence in an English-speaking area) and %USE (daily use of English) variables shown in Table 2-1 and the mean accentedness ratings in Table 4-1. Non-significant correlations of 0.215, 0.179, and -0.003 were obtained.

**Discussion**

The results of this experiment, like those of previous studies, indicate that non-native speakers can be readily distinguished from native English speakers on the basis of pronunciation. In this case, the judges listened only to CVC utterances and were asked to assign accentedness ratings on the basis of how native-like the vowel portions sounded. The fact that the non-native productions received a wide range of scores which were moderately correlated with a number of acoustic predictor variables indicates that the judges were aware of varying degrees of accentedness in the vowels they heard.

A fairly simple approach was used to relate the acoustic properties of the vowels to the accentedness ratings assigned to them. A regression analysis incorporating initial F1 measurements, change in F1 and F2, and target vowel identity accounted for about 43% of the variance in accentedness ratings assigned by the judges. A second model did slightly better in accounting for the rating data, although the amount of improvement obtained could not be adequately assessed.

Although one might try to develop a single regression equation to predict accentedness ratings in any vowel, the evidence presented here suggests that this would be a misguided approach because it appears that the parameters which listeners consider important vary from one vowel to another. While movement in F1 and vowel duration were very potent predictors of accentedness in the vowel /e/ in this experiment, accounting for about 78% of the variance in listeners' judgments, they were less important predictors of the ratings on the other vowels. Rather, the frequency of F1 was the most successful predictor of the ratings given the vowels /ɪ/, /ɛ/, and /æ/. There was some evidence, however, that formant movement played a role in listeners' judgments of /i/, /ɪ/, and /æ/.

The importance of F1 observed here probably reflects, in part, an effect of vowel categorization. Differences in F1 correspond to differences in the traditional dimension of vowel height, the dimension distinguishing the five vowels rated in this experiment. On any given token, a value of F1 outside the range normally encountered in native productions of that vowel may have caused the judges to hear it as belonging to a vowel category other than the one intended. If so, they would most likely assign a low rating to it.

Patterns of formant movement were observed to be closely related to accentedness in the vowel /e/. As observed in Experiment III (Chapter 3) the Arabic subjects generally showed much less movement than the native speakers on this vowel and in several

others. The high correlation between the judges ratings and the $\Delta F1^2$ predictor for this vowel, and the relationship between accentedness and formant movement in some of the other vowels indicates that studies of L2 speakers' vowel productions must assess formant movement if they are to examine thoroughly the characteristics of accented speech which distinguish it from native speech.

Although the results of Experiments I, II, and III reveal rather dramatic differences in the ways native English and native Arabic speakers perceive and produce English vowel duration differences, duration differences emerged as a significant predictor variable in only one vowel examined in Experiment IV. This may indicate that native English listeners do not pay much attention to temporal properties of vowels when assessing accentedness. However, it would be premature to conclude that the often inappropriate vowel durations produced by the non-native speakers did not play some role in accentedness. For instance, the shortness of many of the Arabic speakers' vowels may have resulted in less formant movement than in the native speakers' productions.

While a measurement analysis such as that performed in Experiment III may be useful for determining general differences between native and non-native vowel productions, it does not necessarily reveal which differences will matter most to native speakers assessing accentedness. The data presented here argue rather strongly against the use of simple measurement data as a means of assessing accentedness, unless individual differences are carefully examined. In four of the five front vowels studied in Experiment III no statistically significant differences in F1 frequencies between the two groups emerged, yet the regression analysis in Experiment IV indicated that in two of those four vowels (/ɪ/ and /æ/), *F1* was a significant predictor of accentedness.

Although the analysis presented here may be regarded as successful in the sense that significant predictors of accentedness

ratings were identified, the fact remains that a large part of the variance in the ratings of most vowels was not accounted for. The reasons for this are unclear; obviously, much more study is required to determine how best to refine the approach taken here. One possibility is that the acoustic measurements did not sufficiently characterize the properties of the stimuli which made them sound accented. Perhaps formant measurements at additional times should be taken. Another factor which may have influenced the subjects' judgments was the presentation of vowels in a /bVt/ frame. Although the subjects reported little difficulty in rating only the vocalic portions of the stimuli, they may have been somewhat influenced by consonantal portions if they were noticeably accented. It is also possible that the listeners were influenced by the voice quality of individual speakers. They may have tended, for instance, to assign a certain speaker consistently high or low ratings because his voice was easily recognized.

It should be recognized that the results of any regression analysis depend on the nature of the variability in the predictor variables. Therefore, the results seen here cannot be assumed to reflect all the possible characteristics of non-native vowel productions which might lead listeners to hear English vowels as accented. A lack of a correlation between some predictor and the goodness scores may simply result from a lack of variability in the production data.

While the results of Experiment IV lend some support to the proposal that listeners have access to vowel prototypes against which they can compare individual vowel utterances, much remains unknown about the nature of such prototypes. In fact, the interpretation of the present findings is limited in many respects by the current level of understanding of vowel perception. However, the techniques used here, along with other methods, such as those involving carefully controlled synthetic stimuli, do have considerable

potential in improving our understanding of the nature of vowel categories as well as the phenomenon of foreign accent.

## Tables and Figures for Chapter 4

**Table 4-1: Mean Ratings in Experiment IV by Speaker and Vowel**

| Subj # | /i/ | /ɪ/ | /e/ | /ɛ/ | /æ/ | $\bar{x}$ (SD) |
|---|---|---|---|---|---|---|
| 1 | 77.6 | 64.2 | 79.8 | 13.6 | 35.7 | **54.1 (30)** |
| 2 | 85.3 | 76.3 | 80.3 | 11.5 | 48.6 | **60.4 (30)** |
| 3 | 55.4 | 60.8 | 22.1 | 21.0 | 39.8 | **39.8 (24)** |
| 4 | 62.8 | 19.4 | 77.6 | 72.4 | 44.9 | **55.4 (27)** |
| 5 | 67.7 | 80.6 | 63.3 | 14.7 | 60.0 | **57.3 (25)** |
| 6 | 80.3 | 16.4 | 65.0 | 75.8 | 65.4 | **60.6 (26)** |
| 7 | 71.5 | 33.3 | 33.2 | 31.4 | 45.9 | **43.1 (23)** |
| 8 | 79.5 | 77.0 | 67.4 | 64.3 | 39.3 | **65.5 (20)** |
| 9 | 75.8 | 57.0 | 58.6 | 36.5 | 77.8 | **61.1 (21)** |
| 11 | 74.7 | 34.6 | 78.1 | 42.6 | 53.2 | **56.6 (26)** |
| 13 | 60.2 | 14.7 | 62.0 | 70.3 | 65.9 | **54.6 (23)** |
| 14 | 71.8 | 82.3 | 72.0 | 80.3 | 74.8 | **76.2 (9)** |
| 15 | 60.8 | 34.8 | 24.4 | 43.8 | 64.0 | **45.6 (22)** |
| 16 | 53.7 | 71.3 | 66.2 | 14.6 | 70.2 | **55.2 (25)** |
| 17 | 66.6 | 44.7 | 25.8 | 22.4 | 39.8 | **39.9 (21)** |
| 19 | 82.1 | 75.8 | 38.7 | 14.5 | 75.3 | **57.3 (30)** |
| 20 | 87.9 | 28.2 | 89.3 | 69.0 | 76.4 | **70.2 (25)** |
| 21 | 50.2 | 74.7 | 36.7 | 12.4 | 34.5 | **41.7 (26)** |
| 22 | 84.1 | 79.7 | 88.8 | 62.1 | 76.1 | **78.2 (13)** |
| 23 | 73.8 | 35.2 | 23.3 | 33.7 | 41.1 | **41.4 (24)** |
| 24 | 78.5 | 78.3 | 68.8 | 15.3 | 17.3 | **51.6 (33)** |
| NE1* | 90.9 | 90.3 | 87.6 | 94.8 | 95.0 | **91.7 (8)** |
| NE2* | 79.2 | 88.2 | 84.0 | 92.9 | 84.0 | **85.7 (12)** |
| $\bar{x}$ | **71.4 (11)** | **54.3 (24)** | **58.2 (23)** | **39.1 (25)** | **54.6 (18)** | |

*Native English Speakers. Data not included in means.

**Table 4-2: Inter-rater Agreement in Experiment IV (Pearson r)**

| Judge | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 2 | 0.772 | | | |
| 3 | 0.761 | 0.808 | | |
| 4 | 0.780 | 0.832 | 0.760 | |
| 5 | 0.759 | 0.746 | 0.795 | 0.778 |

**N=115**

**Table 4-3: Predictor Variables in Regression Analysis I**

| Variable | Formula |
|---|---|
| $F1^2$ | $(F1a-F1a_{\bar{x}})^2$ |
| $F2^2$ | $(F2a-F2a_{\bar{x}})^2$ |
| $\Delta F1^2$ | $((F1b-F1a)-(F1b_{\bar{x}}-F1a_{\bar{x}}))^2$ |
| $\Delta F2^2$ | $((F2b-F2a)-(F2b_{\bar{x}}-F2a_{\bar{x}}))^2$ |
| $DUR^2$ | $(Dur-Dur_{\bar{x}})^2$ |

$\bar{x}$=mean value of parameter from EN group data

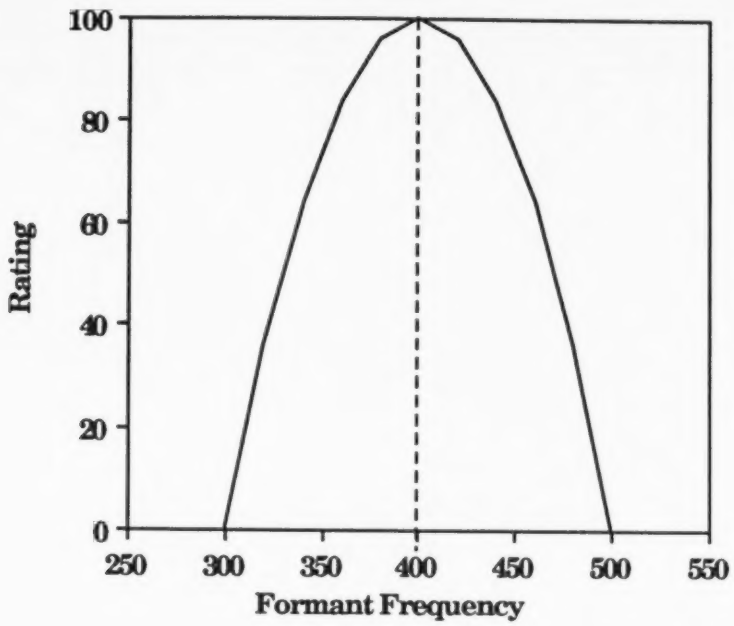
**Table 4-4: Summary of $R^2$ Values from Two Regression Analyses**

| Analysis | All Vowels | /i/ | /ɪ/ | /e/ | /ɛ/ | /æ/ |
|---|---|---|---|---|---|---|
| I | 43 | none | 28 | 78 | 30 | 34 |
| II | 43 | 34 | 38 | 78 | 42 | 57 |

**Table 4-5: Significant Predictors of Judges' Ratings in the Two Analyses**

| Vowel | Model I | Model II |
|---|---|---|
| /i/ | none | $F2$, $\Delta F2^2$ |
| /ɪ/ | $F1^2$ | $F1^2$, $\Delta F2$ |
| /e/ | $\Delta F2^2$, $DUR^2$ | $\Delta F2^2$, $DUR^2$ |
| /ɛ/ | $F1^2$ | $F1$ |
| /æ/ | $F1^2$ | $F1$, $\Delta F1$ |

**Figure 4-1: Hypothetical Rating Function**



**References**

Brennan, E., & Brennan, J. (1981). Measurements of accent and attitude toward Mexican American speech. *Journal of Psycholinguistic Research* **10**:487-501.

Cunningham-Andersson, U., & Engstrand, O. (1989). Perceived strength and identity of foreign accent in Swedish. *Phonetica* **46**: 138–154.

Disner, S. (1986). On describing vowel quality. In J. Ohala and J. Jaeger (Eds.), *Experimental Phonology* (pp 69-79). Orlando: Academic Press.

Flege, J. (1984). The detection of French accent by American listeners. *Journal of the Acoustical Society of America* **76**: 692-707.

Flege, J. (1987). The instrumental study of L2 speech production: some methodological considerations. *Language Learning* **37**: 285-296.

Flege, J. (1988). Factors affecting degree of perceived foreign accent in English sentences. *Journal of the Acoustical Society of America* **84**: 70-79.

Flege, J., & Hillenbrand, J. (1984). Limits on phonetic accuracy in foreign language speech production. *Journal of the Acoustical Society of America* **76**: 708-721.

Flege, J., Munro, M., & Skelton, L. (under review). Production of the word-final English /t/-/d/ contrast by native speakers of Mandarin and Spanish. *Journal of the Acoustical Society of America* .

Greiser, D., & Kuhl, P. (1989). Categorization of speech by infants: support for speech-sound prototypes. *Developmental Psychology* **25**:577-588.

Mitchell, T. (1990). *Pronouncing Arabic I* . Oxford: Clarendon Press.

Munro, M. (1987). *Voicing contrasts in French and English labial stops*. Unpublished M.Sc. Thesis. Edmonton: Department of Linguistics, University of Alberta.

Peterson, G., & Barney, H. (1952). Control methods used in a study of the identification of vowels. *Journal of the Acoustical Society of America* **24**: 175-184.

Purcell, E. & Suter, R. (1980). Predictors of pronunciation accuracy: a reexamination. *Language Learning* **30**: 271-287.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics* **3**: 243-261.

Suter, R. (1976). Predictors of pronunciation accuracy in second language learning. *Language Learning* **26**: 233-253.

Tahta, S., Wood, M., & Loewenthal, K. (1981). Foreign accents: factors relating to transfer of accent from the first language to a second language. *Language and Speech* **24**: 265-272.

# CHAPTER 5
# GENERAL DISCUSSION AND CONCLUSIONS

The research reported here explores, from three different approaches, the perception and production of non-native speech sounds by a group of Arabic-speaking subjects in comparison with a group of native English speakers. In the first study, perceptual experiments examined the role of spectral and temporal cues in the subjects' perceptions of two vowel contrasts, one familiar, and one from a language not spoken by them. In the second study, vowel productions from the two groups were studied with respect to their durations, F1 and F2 frequencies, and F1 and F2 movement. Finally, in the third study, accentedness ratings on five vowels were collected from a panel of native speakers. The use of these three approaches allowed a comparison of the performance of the two groups on perception and production tasks, as well as an examination of the relationship between the acoustic properties of accented speech and accentedness ratings.

## Effects of L1 on L2

In the experiments performed here, it was generally not the case that the Arabic subjects performed in a manner which could be considered native-like. In Experiment I, they showed heavy use of temporal properties relative to spectral ones in their categorizations of the /i/ and /I/ tokens. In Experiment III, it was found that on every vowel the Arabic speakers as a group differed in some respect (e.g. in duration, formant frequency, or formant movement) from the native speakers. Finally, in Experiment IV, all the L2 subjects were rated lower than the native speakers in terms of the goodness of their productions of five English vowels. These differences emerged even though the L2 subjects had been living in the United States for an average of nearly six years. These findings suggest that there are

some aspects of English vowel perception and production which are unlikely to be learned by L2 learners, at least without special training. They are also consistent with the findings of many other studies which indicate that adult L2 learners typically do not develop native-like patterns of perception and production.

One proposal which was considered as an explanation for these results is that the Arabic subjects had not actually learned a second sound system, but rather that they produced and perceived at least some of the sounds of the L2 in terms of their native categories, perhaps with slight modifications. This type of account has been elaborated in different ways by several researchers (Flege, 1987, 1988a; Williams, 1979, 1980; Obler, 1982; Caramazza, Yeni-Komshian, Zurif, & Carbone, 1973). Whether or not this is exactly the case in the present study, there are a number of findings here which support such a view because they demonstrate that the L2 learners were making use of perception and production patterns characteristic of their native language. The inappropriate use of duration on the part of the Arabic-speaking subjects in both the production and perception of English vowels, for instance, shows such an effect. In fact, the evidence collected here confirmed the hypothesis that speakers of a quantity language would show greater perceptual sensitivity to certain vowel duration differences than would speakers of a non-quantity language. It also confirmed reports from other researchers that speakers of Arabic tend to exaggerate duration differences between English tense and lax vowels in production.

Measurements of F1 and F2 frequencies also revealed important differences between the two groups which must in some way be attributable to L1 influences. In the case of /ɛ/, /ʌ/, and /ɑ/, F1 values differed significantly between the native and non-native speakers, while in most of the the back vowels, F2 was lower in the non-native productions. Formant frequency data are not

available from all the dialects of Arabic represented in the sample, so it is not possible to say for certain that these effects correspond to differences which might be readily predicted from the native vowel systems. This, of course, is an issue which could be addressed in future studies if sufficient L1 production data were collected.

The lack of appropriate formant movement which was observed in most of the non-native vowels in the production experiment also indicates an influence of L1. In general, the Arabic speakers showed significantly less movement in F1 and F2 than the English speakers did in vowels in which such movement was expected. Since it is known that Arabic vowels typically show less formant movement than English vowels, the lack of sufficient movement in the non-natives' productions may indicate a lack of perceptual sensitivity to formant movement, or it may again indicate that the speakers were substituting L1 vowels for English ones.

Although this study was not designed as a rigorous test of Flege's (1987, 1988a) proposal concerning the learnability of "new" and "similar" phones, no evidence was found here to support this distinction. On the one hand, it was observed that the Arabic speakers did not perceive or produce a number of similar vowels in a manner which could be considered native-like, as his account predicts. However, the speakers of dialects which do not have the vowel /ɛ/ (i.e., those for whom this vowel is a new vowel) did not appear to perform any better on this vowel than did the speakers of the other dialects.

**Individual Differences**

One observation which has been made several times in the experiments conducted here is that L2 learners show quite dramatic individual differences in their perception and production of sounds from their L2. It was suggested in Chapter 2 that some of this variability may be due to normal variation in how different speakers

make use of different perceptual cues. Even within the monolingual native English group, notable differences were seen in the extent to which the subjects used duration as a cue to the /i/-/ɪ/ distinction and in the degree of duration use in the training task.

There is also evidence here of individual differences in the subjects' success at learning to perceive and produce the English vowels investigated. This was especially clear in Experiment IV, in which it was found that the mean accentedness ratings assigned to the subjects on five vowels varied from a low score of 40 (out of 100), indicating very accented productions, to a high score of 78, which indicated productions that were fairly native-like, although probably still distinguishable from native productions. The data from the other experiments also showed considerable variability of this sort. No evidence was found here that the individual differences were related to the amount of experience the subjects had had with English or with the extent to which they used English in their daily affairs. This is consistent with the view that after an initial period of learning, L2 learners tend to show little additional improvement in their mastery of the L2 sound system, even after several years of experience (Flege, 1988b). For as yet unknown reasons, then, some L2 learners are more successful than others in terms of how native-like their production and perception are. Clearly more research is indicated if we are to understand such differences.

**The Relationship between Perception and Production**

One important issue which has yet to be explored in a detailed, careful way by researchers concerns the relationship between perception and production in the acquisition of the sound system of a second language. Conventional wisdom suggests that L2 learners must learn to perceive a contrast before they learn to produce it, and therefore, that they should not be able to correctly produce contrasts which they do not correctly perceive. However, some researchers

have suggested that this is not the case. Sheldon & Strange (1982), for instance, argued that data from native Japanese learners of English showed that with respect to the English /r/ – /l/ distinction, performance on a production task was actually better than performance on a perception task.

In the present study, an attempt was made to relate the perceptual data from Experiment I to the measurement data and accentedness ratings collected later. Although the relationship between perception and production was explored in several ways here, the evidence that native-like perception is a necessary condition for native-like production was limited. In Experiment III it was observed that some aspects of the production of /i/ and /ɪ/ were correlated with perceptual data on the same two vowels from Experiment I. In particular, subjects who produced the most native-like F2 frequencies for /ɪ/ and showed the most native-like F1 movement for /i/, tended to perform more like the native speakers on the perception task. However, no connection was found between the perceptual data and the duration data from Experiment III. Nor was there a relationship with the accentedness judgments from Experiment IV. In other words, the subjects who heard the /i/–/ɪ/ contrast in a native-like way did not necessarily receive higher scores on the rating task.

Perhaps part of the reason for the lack of correspondence between the perception and production data presented here is simply that the ability to perceive a contrast correctly does not necessarily guarantee the ability to produce it correctly. For instance, it seems quite reasonable to suppose that poor production may occur because a speaker is unable to implement knowledge gained through perception. Therefore, correlational analyses such as those performed here may fail to yield positive results.

There is good reason for caution, however, in drawing conclusions about studies which have suggested that good production

can develop in the absence of good perception. In the Sheldon & Strange study, data on the English /r/-/l/ contrast were collected from six Japanese speakers. Success on the production study was determined by correct identification scores from a panel of native English listeners, while the perceptual data were actually scores from an identification task performed by the Japanese speakers on native productions. As mentioned previously, Sheldon & Strange reported higher scores on the production task than on the perception task. However, a comparison of two sets of data such as these can lead to some overly simplistic conclusions. For instance, even though the Japanese speakers' productions were well–identified by the English speakers, it does not follow that they were always very good exemplars of English /r/ and /l/. No assessment of goodness was elicited from the English listeners. Second, the Japanese subjects actually performed quite well on the perceptual task, with all but one of them scoring over 80% correct and three of the six scoring above 92%. Clearly these speakers could have had a strong perceptual basis for their productions. Third, the results do not prove that production of a particular contrast can *precede* perception of it, as they conclude, because their study did not compare speakers at different stages of development. Rather, it examined both production and perception at a single time in a single group of subjects. The results therefore do not reveal how the L2 learners' perception and production actually developed. Finally, their results cannot be taken to mean that the ability to produce a contrast *normally* develops without the ability to perceive it.

Gass (1984) seems to have made the latter assumption in her study of VOT in the perception and production of English /p/ and /b/ by speakers of several languages. She suggests that because her subjects labelled a synthetic VOT continuum in a rather different way from native speakers (because identification functions were generally less steep), but did produce VOT values close to native

norms, the subjects' perceptual abilities lagged their production abilities. Again, such an interpretation does not clearly follow from the data. In fact, there is no reason to suppose that L2 learners need to identify stimuli in a perfectly native-like way in order to use knowledge gained through perception to produce it well enough for native speakers to correctly identify their productions, perhaps with perfect accuracy. Therefore, Sheldon & Strange (1982) and Gass (1984) do not convincingly demonstrate that production can "lag" perception. Furthermore, the statement by Sheldon & Strange that "perceptual and productive mastery of the /r/-/l/ contrast was shown here to be independent of each other (p.254)" is doubtful. As they suggest, it might be possible that with training in which learners are given specific articulatory instructions on how to produce a particular L2 sound, fairly accurate production may be achieved in the absence of a corresponding perceptual ability. But this is rather different from the suggestion that perception and production can be independent.

The relationship between perception and production is an important one for second language pedagogy as well as for researchers developing methods for training subjects to produce and perceive non-native sounds. It is necessary, then, that the issues raised here be explored further. The results of this study give some encouragement that relationships between perception and production might emerge in correlational analyses, but the connection still appears to be rather elusive.

**Measuring Accentedness**

Experiments III and IV demonstrate a clear relationship between certain acoustic properties of accented vowels and ratings assigned to them by native speakers. They indicate that accentedness in vowels is to some extent "measurable." The evidence presented here suggests that the properties which make vowels sound accented

may differ from vowel to vowel. While formant movement was important in /e/, F1 frequency was more important in other front vowels.

Studies such as Experiment IV may prove useful in improving our general understanding of the nature of phonetic representations. Accented productions are particularly useful in this regard because they can be used to explore listeners' perceptions of goodness in both vowel and consonant categories using natural tokens. The findings reported here lend some support to the proposal that listeners have access to category prototypes against which they may compare individual exemplars when they assign accentedness ratings.

In conclusion, the present study has demonstrated a number of techniques for studying foreign accent in L2 learners. The results indicate that the adult L2 learners here differed considerably from native speakers both in their perception and in their production of a number of English vowels, although important individual differences emerged in every experiment. The fact that many of the errors in both perception and production observed here were attributable to characteristics of the subjects' first language indicates that foreign accent is to some extent a predictable phenomenon. Moreover, the results of the accentedness rating study indicate that foreign accent can be quantified in terms of acoustic measurements which show a systematic relationship with native listeners' perceptions of goodness.

**References**

Caramazza, A., Yeni–Komshian, G., Zurif, E., & Carbone, E. (1973). The acquisition of a new phonological contrast: the case of stop consonants in French-English bilinguals. Journal of the Acoustical Society of America 54: 421-428.

Flege, J. (1987). The production of "new" and "similar" phones in a foreign language: evidence for the effect of equivalence classification. *Journal of Phonetics* **15**: 47-65.

Flege, J. (1988a). The production and perception of foreign languages. In Winitz, H. (Ed.). *Human communication and its disorders* (pp. 224-401). Norwood, NJ: Ablex.

Flege, J. (1988b). Factors affecting degree of perceived foreign accent in English sentences. *Journal of the Acoustical Society of America* **84**: 70-79.

Gass, S. (1984). Development of speech perception and speech production abilities in adult second language learners. *Applied Psycholinguistics* **5**: 51-74.

Obler, L. (1982). The parsimonious bilingual. In Obler, L., & Menn, L. (Eds.), *Exceptional Language and Linguistics* (pp. 339-346). New York:Academic Press.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics* **3**: 243-261.

Williams, L. (1979). The modification of speech perception and production in second–language learning. *Perception & Psychophysics* **26**: 95-104.

Williams, L. (1980). Phonetic variation as a function of second-language learning. In G. Yeni-Komshian, J. Kavanagh, & C. Ferguson (Eds.) *Child phonology, Vol. 2: Perception* (pp. 185-215). New York: Academic Press.

# APPENDICES

## Appendix A: Instructions to Subjects

### Instructions for Experiment I

You will hear a number of words played through the headphones. These words were synthetically produced; that is, they were generated on a computer. In each case, listen carefully to the word, and decide whether it sounds more like the English word 'beat' (as in 'Beat it!') or the English word 'bit' (as in 'The dog bit me.') Then press the button labelled with the correct word. The computer will wait for you to press the button before it plays the next word. If you are unsure which word was played, make your best guess. Do you have any questions?

### Instructions for Experiment II

The purpose of this experiment is to find out how people classify vowels from a language they don't know. You will hear several examples of two different words which do not occur in English (or Arabic). Since they are foreign words we will not try to spell them. Instead we will use colours to represent them. One will be the 'blue' word and the other the 'orange' word. These words sound similar, but they have different vowel sounds in them. Your task is to figure out which are blue ones and which are orange ones.

At first you will simply have to guess which is which, and the computer will tell you whether you have guessed correctly. For example, if you think the word is a blue word, press the button next to the blue square. If you are correct, the small yellow light near the blue square will come on. If you are wrong, the yellow light near the

orange square will come on. Very soon you will be able to hear the difference between the vowels in the words. It is important to listen carefully to the vowel portion before making your decision.

The experiment will continue for several minutes. Occasionally there may be a short pause, during which you can take a rest. After a while, the computer will stop telling you whether your response is correct or not. When this happens, please keep pressing the buttons just as before.

This task takes about 10 to 15 minutes. Do you have any questions?

## Appendix B: Nominal and Measured Properties of Stimuli in Experiment I

### Formant Frequencies and Bandwidths of Endpoint Stimuli (Hz)

| | Extreme /i/ | | Extreme /ɪ/ | |
|---|---|---|---|---|
| **Par.** | **Nom.** | **Measured** | **Nom.** | **Measured** |
| F1 | 233 | 230 (0 dB) | 361 | 352 (0 dB) |
| F2 | 2400 | 2409 (-19 dB) | 2000 | 2009 (-15 dB) |
| F3 | 303 | 3090 (-21 dB) | 2760 | 2759 (-20 dB) |
| F4 | 3850 | 3712 (-25 dB) | 3850 | 3798 (-22 dB) |
| B1 | 90 | 56 | 90 | 53 |
| B2 | 100 | 69 | 100 | 84 |
| B3 | 300 | 216 | 300 | 242 |
| B4 | 500 | 532 | 500 | 352 |

### Measured Durations of Stimuli at Spectral Step 1 (ms)

| Dur. Step | Nominal | Measured |
|---|---|---|
| 11 | 250 | 253.3 |
| 12 | 225 | 226.4 |
| 13 | 200 | 208.3 |
| 14 | 175 | 180.9 |
| 15 | 150 | 153.9 |
| 16 | 125 | 126.6 |

**Appendix C: Nominal and Measured Properties of Stimuli in Experiment II (Hz)**

| Token | Parameter | Nominal | Measured |
|---|---|---|---|
| 1. /byt/ | F1 | 297 | 272 |
| | F2 | 1828 | 1828 |
| | F3 | 2137 | 2125 |
| 2. /byt/ | F1 | 314 | 305 |
| | F2 | 1719 | 1722 |
| | F3 | 2137 | 2125 |
| 3. /byt/ | F1 | 349 | 348 |
| | F2 | 1866 | 1884 |
| | F3 | 2137 | 2102 |
| 4. /byt/ | F1 | 280 | 247 |
| | F2 | 1942 | 1942 |
| | F3 | 2137 | 2109 |
| 5. /byt/ | F1 | 247 | 233 |
| | F2 | 1791 | 1798 |
| | F3 | 2137 | 2127 |
| 6. &10. /byt/ | F1 | 346 | 346 |
| | F2 | 1780 | 1791 |
| | F3 | 2137 | 2124 |
| 7. & 11. /byt/ | F1 | 322 | 316 |
| | F2 | 1934 | 1934 |
| | F3 | 2137 | 2110 |
| 8. &12. /byt/ | F1 | 249 | 234 |
| | F2 | 1878 | 1899 |
| | F3 | 2137 | 2100 |
| 9. & 13. /byt/ | F1 | 272 | 243 |
| | F2 | 1726 | 1735 |
| | F3 | 2137 | 2126 |
| 14. /bøt/ | F1 | 366 | 356 |
| | F2 | 1462 | 1429 |
| | F3 | 2290 | 2315 |
| 15. /bøt/ | F1 | 384 | 355 |
| | F2 | 1326 | 1321 |
| | F3 | 2290 | 2442 |
| 16. /bøt/ | F1 | 421 | 401 |
| | F2 | 1452 | 1442 |
| | F3 | 2290 | 2441 |
| 17. /bøt/ | F1 | 348 | 344 |
| | F2 | 1518 | 1516 |
| | F3 | 2290 | 2445 |
| 18. /bøt/ | F1 | 313 | 295 |
| | F2 | 1388 | 1394 |
| | F3 | 2290 | 2446 |

**Appendix C (continued)**

| | | | |
|---|---|---|---|
| 19. & 23. /bøt/ | F1 | 418 | 391 |
| | F2 | 1378 | 1382 |
| | F3 | 2290 | 2444 |
| 20. & 24. /bøt/ | F1 | 392 | 360 |
| | F2 | 1511 | 1489 |
| | F3 | 2290 | 2449 |
| 21. & 25. /bøt/ | F1 | 316 | 304 |
| | F2 | 1462 | 1449 |
| | F3 | 2290 | 2442 |
| 22. & 26. /bøt/ | F1 | 340 | 335 |
| | F2 | 1333 | 1326 |
| | F3 | 2290 | 2441 |

The measured duration of the vowels in the long stimuli was about 205 ms (nominal 200 ms). The vowels in the short stimuli were about 127 ms long (nominal 125).